

Projet Deep Learning - Big Data MAPI3 2024

Léo Andéol
leo.andéol@math.univ-toulouse.fr

24 Janvier 2024

Ce projet constitue la deuxième partie du projet BigData.

Vous devrez, pour l'évaluation, rendre un notebook python détaillé avec des commentaires justifiant les méthodes utilisées.

Vous devrez aussi présenter vos résultats et les méthodes en 20 minutes lors de la soutenance du 07/02

La performance de vos modèles ne sera pas utilisée comme critère d'évaluation. Nous nous intéresserons uniquement à la méthodologie utilisée, aux comparaisons et analyses produites, ainsi qu'à la variété d'expériences menées. Vous pouvez faire la présentation et le notebook en anglais si vous préférez.

Vous êtes encouragés à :

- utiliser les notebooks de TP comme base de travail.
- utiliser Google Colab pour l'entraînement de vos modèles.
- utiliser la bibliothèque keras pour l'entraînement (tensorflow et pytorch sont acceptés).

Vous devez rendre votre notebook avant le 06/02 à 23h59. Vous devez nous le transmettre par mail à

- leo.andéol@math.univ-toulouse.fr,
- francois.malgouyres@math.univ-toulouse.fr et
- francois.bachoc@math.univ-toulouse.fr

en précisant dans l'objet du mail "Projet Deep Learning - Big Data MAPI3 2024".

1 Sujet :

Vous allez, dans ce projet, entraîner plusieurs réseaux de neurones à classifier des paysages. Vous utiliserez pour cela le dataset Landscapes disponible ici.

Il s'agit d'un dataset comprenant 4000 images appartenant à 7 catégories.

Vous devez :

1. Séparer vos données en 2 : un jeu d'apprentissage 80% et un jeu de validation 20%
2. Entraîner un réseau de convolutions que vous aurez défini vous même



FIGURE 1 – Exemples du dataset

3. Utiliser de manière individuelle différentes techniques permettant d'augmenter la capacité de généralisation afin d'améliorer votre modèle (data augmentation, dropout, etc...)
4. Utiliser un réseau pré-entraîné pour améliorer vos performances
5. Comparer les résultats produits par 2. 3. et 4. dans un tableau récapitulatif (utilisez le jeu de validation).
6. Afficher quelques exemples sur lesquels votre réseau se trompe et afficher ses prédictions
7. Visualiser vos projections dans l'espace des features à l'aide d'une t-SNE sur un sous ensemble du dataset (quelques centaines d'images).
8. A l'aide des features calculées par votre réseau, choisir une image au hasard dans le dataset et afficher les 3 images qui sont les plus proches au sens de la distance euclidienne. Vous pourrez, si vous le souhaitez, utiliser les plus proches voisins de Scikit-Learn