La thématique choisie pour ce projet est le Vision Transformers. L'objectif principal est de se familiariser avec une méthode qui représente l'état de l'art en matière de computer vision et de comparer ses résultats avec des méthodes classiques telles que les réseaux convolutionnels. Il sera ainsi possible de juger de la plus-value que représentent les vision transformers.

En ce qui concerne le jeu de données, mon choix s'est porté sur Stanford Cars en raison de sa similarité avec Stanford Dogs (196 classes pour le premier versus 120 pour le second) que j'ai déjà traité le projet précédent.

La baseline est un réseau convolutionnel classique composé de 10 layers. Quant à la méthode état de l'art, ce sera un modèle vision transformer pré entraîné sur la banque ImageNet.

Le plan de travail est le suivant :

- Extraire un train set de 5 classes du jeu de données originel, puis l'entraîner sur la baseline. Les hyperparamètres de ce dernier seront fine-tunés.
- Entraîner le modèle fine-tuné sur l'ensemble du train set et l'évaluer sur le test set.
- Déployer le modèle vision transformer. L'évaluer sur un jeu de validation puis sur le jeu de test.

Bibliographie :

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei.
3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov,Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner,Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby.
An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

https://towardsdatascience.com/getting-started-with-pytorch-image-models-timm-a-practitioners-guide-4e77b4bf9055#5e2d