Développer une preuve de concept

Plan de travail prévisionnel

Problématique:

Lors d'un précédent projet, nous avions utilisé le réseau de neurones convolutif Xcpetion pour procéder à la classification d'images de chiens en fonction de la race. Les réseaux de neurones convolutifs, ou CNN, sont actuellement les méthodes les plus en vues dans le domaine de la vision par ordinateur. Cependant, depuis quelques mois, une nouvelle méthode vient concurrencer les CNN. Il s'agit du Vision Transformer, ou ViT, une adaptation au traitement d'images des Transformers couramment utilisé en traitement du langage.

L'idée est donc de comparer, en termes de précision et temps de calcul, un CNN et un ViT. Pour ce faire, nous réutiliserons le Stanford Dogs Dataset et le modèle Xception précédemment entrainé que l'on comparera au modèle ViT-B/16 de Google.

Méthodologie:

- Nous travaillerons avec le data set Stanford Dogs Dataset (Stanford Dogs dataset for Fine-Grained Visual Categorization)
- Nous utiliserons le modèle Xception comme méthode de référence
- Nous entraînerons le modèle ViT-B/16 en utilisant une méthode de fine-tuning
- La comparaison des modèles se fera à l'aide du temps de calcul et du score de précision (Accuracy)

Sources bibliographiques:

- Article de recherche : 2010.11929v2.pdf (arxiv.org)
- Article de vulgarisation : Vision Transformers (ViT) in Image Recognition - 2022 Guide - viso.ai
- Code source: google-research/vision_transformer (github.com)
- Tutoriel : Fine-Tune ViT for Image Classification with 🤗 Transformers (huggingface.co)