DataSpace

Plan prévisionnel

Dataset retenu

Le dataset est le même que pour le projet précédent : Classez des images à l'aide d'algorithmes de Deep Learning. Pour l'entraînement des modèles personnalisés seulement 10 races ont été gardées, les 120 races initiales produisent des résultats médiocres. Les 10 races comportant 1'865 images.

Modèle envisagé

Dans le cadre du projet précédent : Classez des images à l'aide d'algorithmes de Deep Learning, un modèle réseaux de neurones convolutifs (modèle CNN) a été entraîné afin de pouvoir catégoriser des images de chiens en fonction de leur race.

Après avoir utilisé des transformers avec succès dans le domaine du langage naturel, ces derniers ont été appliqués au domaine de la vision par ordinateur, c'est ainsi que l'architecture VIT (Vision Transformers) est née. Elle permet d'obtenir une haute précision malgré un temps d'apprentissage considérablement réduit par rapport aux réseaux CNN courants, comme démontré dans le document de recherche : AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE.

Un nouveau modèle VIT personnalisé va être entraîné dans le cadre de ce projet et sera comparé au modèle CNN existant.

Références bibliographiques

« Attention Is All You Need », 2017.

https://arxiv.org/abs/1706.03762

Auteurs : Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin.

Ce document présente l'architecture de transformers qui utilise le mécanisme d'attention. Dans le domaine du langage naturel ce dernier permet de pondérer la pertinence des mots dans une phrase pour obtenir de meilleurs résultats plus rapides pour par exemple une traduction.

« An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale », 2020.

https://arxiv.org/abs/2010.11929

Auteurs : Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby.

Ce document présente l'utilisation des transformers pour la classification d'images. Il introduit également la notion de diviser l'image en multiple patches/segments pour réduire l'utilisation des ressources et d'accélérer l'entraînement du modèle.

Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)

Un modèle CNN personnalisé pour classifier les races de chiens a déjà été produit dans le projet précédent. Un nouveau modèle personnalisé VIT va être entraîné sur le même set de données. Ces modèles seront ensuite comparés du point de vue de leurs pertes, précision et temps d'entraînement.

Une application Streamlit va permettre de présenter le set de données et d'effectuer une prédiction sur une image téléversée en utilisant le nouveau modèle VIT.