Plan prévisionnel

Dataset retenu

*Le dataset **Stanford Dogs** est composé d'images de chiens, réparties sur 120 races différentes, ce qui en fait un ensemble de données riche et varié pour des tâches de classification fine. Ce dataset inclut plus de 20 000 images étiquetées, offrant une grande diversité de poses, d'arrière-plans, et de conditions d'éclairage. C'est un excellent choix pour comparer l'efficacité de différents modèles de classification d'images, car les différences entre races sont parfois subtiles, mettant ainsi à l'épreuve la capacité des modèles à capturer des détails spécifiques et à distinguer des traits visuels fins.*

Modèle envisagé

Pour cette preuve de concept, le Vision Transformer (ViT) a été retenu pour sa capacité à capturer efficacement les relations globales au sein des images, ce qui le rend particulièrement adapté à des tâches de classification où la structure spatiale des données est complexe. Contrairement aux CNN traditionnels, qui se concentrent principalement sur des relations locales via des convolutions, les Transformers, initialement utilisés dans le traitement du langage naturel, exploitent des mécanismes d'attention pour établir des connexions longues distances entre différentes parties de l'image. Le Vision Transformer a montré des performances compétitives et parfois supérieures aux CNN classiques, en particulier sur des tâches de reconnaissance d'images de haute dimension, et s'avère prometteur pour les classifications complexes comme celles de races de chiens.

Références bibliographiques

Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv. Cet article introduit le Vision Transformer (ViT) et explore comment ce modèle, initialement conçu pour le NLP, peut surpasser les CNN en termes de performances de classification d'images.

Blog de Machine Learning Mastery. What is the Vision Transformer (ViT), and Why Does It Matter? Cet article explique en détails le fonctionnement des Transformers pour la vision et leurs avantages par rapport aux architectures CNN classiques, en rendant accessible la compréhension des mécanismes d'attention et des patchs.

Kumar, R., et al. (2021). Comparative Analysis of CNNs and Vision Transformers on Image Classification Tasks. Un article de recherche qui analyse les performances de ViT face aux CNN sur plusieurs ensembles de données, démontrant les cas où ViT est plus performant, notamment sur les tâches de classification d'images fines et variées.

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

Baseline : Utilisation des performances optimales du VGG sur le dataset Stanford Dogs, avec les hyperparamètres déjà optimisés pour obtenir un benchmark clair.

Nouvel algorithme (ViT) : Entraînement du Vision Transformer sur le même dataset, avec un découpage des images en patches et un ajustement des hyperparamètres clés tels que le taux d'apprentissage et la taille des patches.

Comparaison des performances : Évaluation des deux modèles en utilisant des métriques de précision, de recall, et de F1-score, ainsi qu'une analyse du temps d'entraînement et des ressources nécessaires pour chaque modèle.

Interface et visualisation : Création d'un tableau de bord interactif via Streamlit pour illustrer les résultats comparés entre ViT et CNN, avec des visualisations des images mal classées pour analyser les différences de traitement et les cas où chaque modèle excelle ou échoue.