**Note méthodologique : preuve de concept**

## **Dataset retenu**

*Le dataset Stanford Dogs est spécifiquement conçu pour des tâches de classification d’images fines en vision par ordinateur. Composé de 20 580 images réparties en 120 races de chiens, cet ensemble de données est l’un des benchmarks les plus populaires pour la reconnaissance de races de chiens. Sa richesse et sa complexité en font un choix idéal pour évaluer la capacité des modèles à distinguer des catégories très proches visuellement.*

*Structure du dataset : Les images sont divisées en un ensemble d'entraînement et un ensemble de test, ce qui facilite l’évaluation des modèles. Chaque image est annotée avec une étiquette correspondant à la race du chien, et les images présentent une grande diversité de poses, d’angles, de conditions d’éclairage et de contextes, ce qui pose un défi supplémentaire pour les algorithmes de classification d’images. Cette variabilité permet de simuler des conditions réelles de reconnaissance visuelle, rendant les performances des modèles plus représentatives de leur robustesse en production.*

*Pourquoi le Stanford Dogs ? Le choix du Stanford Dogs est justifié par :*

*La précision requise pour différencier les races : Les différences entre races sont souvent subtiles (textures du pelage, formes de la tête, tailles), testant les limites de résolution des modèles et leur capacité à capturer des détails visuels fins.*

*Un benchmark populaire : Le dataset est fréquemment utilisé dans la littérature pour évaluer des modèles de classification d’images. Cela permet de comparer directement les résultats obtenus avec des standards de l’industrie et de la recherche.*

*Pertinence pour une comparaison Transformer vs CNN : La structure du dataset, avec des classes très proches les unes des autres, est idéale pour explorer les différences de performance entre les modèles de type CNN, qui se concentrent sur des relations locales, et le Vision Transformer (ViT), qui utilise des relations globales.*

## **Les concepts de l’algorithme récent**

*Le Vision Transformer (ViT), introduit par Dosovitskiy et al. en 2020, est une adaptation des Transformers, initialement conçus pour le traitement de texte, aux tâches de vision par ordinateur. Ce modèle représente une avancée dans le domaine de la classification d’images en se démarquant des architectures CNN traditionnelles grâce à sa capacité à capturer des relations globales dans les images.*

*Principe de fonctionnement du Vision Transformer*

*Les Transformers sont basés sur un mécanisme d'attention qui permet de modéliser efficacement les dépendances entre différents éléments d'entrée. Dans les tâches de traitement du langage naturel, les Transformers se révèlent capables de capter des relations longues distances entre les mots, une qualité qui peut être transposée au domaine de la vision pour modéliser les relations entre différentes régions d’une image.*

*Découpage de l'image en patches : Contrairement aux CNN qui utilisent des convolutions pour extraire des caractéristiques à différentes échelles de l’image, le ViT divise chaque image en une série de petits carrés, ou "patches" (par exemple, de 16x16 pixels). Chaque patch est ensuite linéarisé et projeté dans un espace de caractéristiques (embedding) avant d'être traité comme un "mot" dans une séquence. Ce processus simule le traitement des mots dans un texte, permettant au Transformer de traiter l’image comme une séquence de patches.*

*Encodage de position : Dans le traitement du texte, l’ordre des mots est fondamental. Cependant, les Transformers n'intègrent pas naturellement l'ordre des éléments de séquence. Pour pallier ce manque, un encodage de position est ajouté à chaque patch, permettant au modèle de conserver l’information de position spatiale de chaque segment de l’image. Cela permet à ViT de distinguer les différentes régions de l’image et de capter les relations spatiales nécessaires pour l’interprétation visuelle.*

*Mécanisme d’attention et couches de Transformer : Les patches linéarisés et encodés en position passent ensuite par plusieurs couches de self-attention. Le mécanisme d’attention calcule des poids pour chaque relation entre les patches, mettant en valeur les parties de l’image les plus importantes pour la tâche de classification. Cette capacité à pondérer les relations permet à ViT d’identifier des structures globales dans l’image, en capturant des dépendances à longue distance et en s’adaptant ainsi aux images complexes comme celles du dataset Stanford Dogs.*

*Classification via un token de classification : Après le passage par plusieurs couches d’attention, le Vision Transformer utilise un token de classification (ou "class token") qui recueille les informations globales extraites de l’image. Ce token est ensuite utilisé pour prédire la classe de l’image, en combinant toutes les informations contextuelles extraites des patches par le mécanisme d’attention.*

*Contexte d’utilisation et spécificités du Vision Transformer*

*Les ViTs se révèlent particulièrement performants dans des contextes où :*

*Les relations globales dans l'image sont cruciales : Contrairement aux CNN qui se concentrent principalement sur les relations locales, ViT excelle dans les situations où il est essentiel de capter des relations de longue distance entre différentes parties de l'image. Par exemple, pour distinguer entre des races de chiens, certaines caractéristiques globales (comme la forme générale du corps ou la position des oreilles) peuvent être plus importantes que des détails locaux.*

*Des données en grandes quantités sont disponibles : Les ViTs nécessitent généralement des ensembles de données de grande taille pour être performants, car ils n’intègrent pas de biais inductifs comme les CNN. Le Stanford Dogs dataset, bien que plus restreint que les grands benchmarks, permet tout de même d’entraîner le ViT avec un certain niveau de performance.*

*Les avantages du Vision Transformer par rapport aux CNN*

*Captation de relations globales : ViT surpasse souvent les CNN sur des tâches nécessitant une vision globale de l'image, car il n’est pas limité par des fenêtres locales de convolution.*

*Efficacité computationnelle sur des patches : En divisant l'image en patches, ViT réduit la complexité de traitement, ce qui permet d'entraîner le modèle de manière plus rapide et plus flexible sur des architectures modernes.*

*Flexibilité d'échelle : Avec des versions telles que ViT-Base, ViT-Large, et ViT-Huge, le modèle peut être ajusté en fonction des ressources disponibles, permettant une optimisation entre précision et temps de calcul.*

*Dans le cadre de ce projet, le Vision Transformer sera comparé aux CNN traditionnels pour tester sa capacité à réduire le nombre de faux positifs, en exploitant sa capacité à capter des relations complexes entre les différentes régions d'une image, même lorsqu'il s'agit de races de chiens présentant des similarités visuelles.*

**La modélisation**

*La méthodologie de modélisation vise à comparer les performances du Vision Transformer (ViT) aux CNN traditionnels en évaluant leur capacité à classer correctement les images du dataset Stanford Dogs, avec un focus particulier sur les faux positifs comme métrique d’évaluation*

*Métrique d’évaluation : Faux Positifs*

*Les faux positifs représentent ici les erreurs où le modèle attribue une race incorrecte à une image. Dans un contexte de classification d'images, cette métrique est particulièrement pertinente car elle permet de mesurer la capacité du modèle à éviter des erreurs de classification qui pourraient être confondues visuellement. Minimiser les faux positifs est crucial pour obtenir un modèle fiable dans la différenciation des races de chiens, surtout lorsque certaines races présentent des caractéristiques physiques proches.*

*Méthodologie de modélisation*

*La modélisation s’effectuera en deux phases distinctes, avec une baseline basée sur un CNN optimisé (VGG), suivie de la modélisation avec le Vision Transformer.*

*Baseline avec CNN (VGG)*

*La baseline repose sur un modèle CNN préentraîné, en particulier VGG, dont les hyperparamètres ont déjà été optimisés pour ce dataset. VGG est un modèle efficace pour la classification d’images, connu pour ses couches convolutives profondes qui capturent des caractéristiques locales dans les images. Ce modèle servira de référence pour évaluer les performances du Vision Transformer.*

*Architecture et hyperparamètres : Le modèle VGG sera configuré avec les hyperparamètres optimaux déjà établis (par exemple, learning rate, nombre d’époques, taille de batch).*

*Entraînement : Le modèle sera entraîné sur l’ensemble d’entraînement du dataset Stanford Dogs et évalué sur un ensemble de test, avec une attention particulière aux faux positifs dans les résultats.*

*Nouvel algorithme avec Vision Transformer (ViT)*

*La deuxième phase implique l’entraînement du Vision Transformer, dont l'architecture sera ajustée pour correspondre au format et aux dimensions du dataset. ViT, contrairement aux CNN, fonctionne sur des patches linéarisés de l’image, capturant des informations globales.*

*Prétraitement des images : Les images du dataset sont redimensionnées (par exemple, à 255x255 pixels) puis divisées en patches de 16x16 pixels, chacune étant ensuite projetée dans un espace de caractéristiques.*

*Entraînement et hyperparamètres : Les principaux hyperparamètres (learning rate, taille de patch, nombre de têtes d’attention) seront ajustés pour maximiser la performance de ViT. Le modèle sera entraîné avec des taux de dropout adaptés pour éviter le surapprentissage, particulièrement important pour le ViT qui peut être sujet à un surajustement si le nombre de données est limité.*

*Évaluation : Comme pour la baseline, le ViT sera évalué sur l’ensemble de test, en mettant l’accent sur la réduction des faux positifs.*

*Optimisation des Modèles*

*Les deux modèles (VGG et ViT) feront l’objet de quelques ajustements pour maximiser leurs performances, en optimisant des hyperparamètres spécifiques :*

*CNN (VGG) : En raison des optimisations déjà existantes, les ajustements seront minimaux. Cependant, des régulations additionnelles, comme une légère augmentation de dropout, pourraient être appliquées pour éviter le surajustement.*

*Vision Transformer : Étant plus sensible aux variations d’hyperparamètres, plusieurs ajustements seront testés, comme :*

*Taille des patches : L'impact de la taille de patch (par exemple, 16x16 vs 32x32) sera exploré pour observer son effet sur les faux positifs.*

*Nombre de couches et de têtes d’attention : ViT utilise des couches empilées de self-attention ; augmenter ou diminuer le nombre de têtes d’attention peut améliorer la capacité du modèle à capturer les relations dans les images.*

*Dropout : Un taux de dropout ajusté est essentiel pour ViT, notamment pour éviter le surajustement dû à sa complexité.*

*Comparaison et analyse des performances*

*La comparaison se concentrera principalement sur les faux positifs, afin de déterminer si le ViT réduit mieux ces erreurs que le VGG. Une analyse détaillée des erreurs sera également effectuée pour comprendre dans quelles situations ViT ou VGG commettent plus de faux positifs. Cette comparaison permettra de valider l’hypothèse selon laquelle le Vision Transformer peut capturer des relations plus globales dans l’image, ce qui pourrait être crucial pour différencier les races de chiens qui partagent des traits visuels similaires.*

**Une synthèse des résultats**

*L’analyse des performances du Vision Transformer (ViT) par rapport au modèle baseline CNN (VGG) révèle des différences marquées dans la capacité de chaque modèle à minimiser les faux positifs. Cette synthèse présente les résultats comparés en termes de précision globale, de taux de faux positifs, et de temps de calcul, suivie d'une conclusion sur les performances de ces modèles pour la classification des races de chiens.*

*Résultats comparés entre ViT et CNN (VGG)*

*Précision et Taux de Faux Positifs*

*Précision globale : Le Vision Transformer atteint une précision bien supérieure à celle du VGG sur l'ensemble de test, et ce plus rapidement (nombre d’époques nécessaires moindre par rapport à VGG) pour un taux d’apprentissage de l’optimiseur égal à celui utilisé pour VGG et en particulier pour les races de chiens qui partagent des caractéristiques physiques communes.*

*Taux de Faux Positifs : Le ViT démontre une grosse réduction du taux de faux positifs par rapport au VGG, confirmant son avantage pour éviter des confusions entre races proches. L’attention globale capturée par ViT semble permettre une meilleure distinction entre des races visuellement similaires, contribuant à des erreurs moins fréquentes en termes de classification incorrecte.*

*Analyse des erreurs*

*Les faux positifs subsistent dans les deux modèles sont notamment pour des images en angles ou en arrière-plans difficiles. Ces cas suggèrent des limitations potentielles des deux architectures en matière de généralisation sur des images moins typiques.*

*Efficacité de l’Entraînement et du Temps de Prédiction*

*Temps d'entraînement : En raison de sa complexité, ViT nécessite un temps d'entraînement légèrement supérieur à celui de VGG, surtout pour des patches et des couches d’attention optimisés.*

*Temps de prédiction : Bien que le ViT demande plus de ressources en calcul, l’écart reste acceptable et viable dans un environnement de production, surtout compte tenu de la réduction des faux positifs obtenue.*

*Conclusion*

*L'expérimentation montre que le Vision Transformer (ViT) dépasse le VGG en termes de réduction des faux positifs, notamment grâce à son approche par attention, qui exploite les relations globales entre les différentes parties de l’image. Ce résultat valide l'hypothèse initiale selon laquelle les Transformers peuvent apporter une valeur ajoutée par rapport aux CNN traditionnels dans les tâches de classification fine, où des détails globaux sont essentiels pour distinguer les catégories proches.*

*En conclusion, le Vision Transformer s'avère prometteur pour des applications en vision par ordinateur nécessitant une haute précision, et notamment pour la classification d’images complexes. Les résultats obtenus ouvrent des perspectives pour une adoption plus large des Transformers dans des projets nécessitant une reconnaissance visuelle précise, avec des possibilités d'amélioration via l’ajustement de la taille de patch et du nombre de couches d’attention.*

**L’analyse de la feature importance globale et locale du nouveau modèle**

*Analyse des métriques d’évaluation pour le Visual Transformer*

*Contrairement aux approches basées sur les réseaux de neurones convolutifs (CNN), pour lesquelles l’interprétation des pixels ou des caractéristiques locales peut être réalisée à l’aide d’outils comme l’analyse de la feature importance, l’évaluation de la pertinence d’un modèle Visual Transformer repose principalement sur l’analyse des métriques d’évaluation. Ces métriques fournissent une compréhension directe des performances du modèle, notamment en termes de classification et de gestion des erreurs.*

*Importance des métriques d’évaluation pour le Visual Transformer*

*L’évaluation de la précision et des erreurs (notamment les faux positifs) est cruciale pour déterminer l’efficacité du Visual Transformer sur le Stanford Dogs dataset. Étant donné la complexité des images et la diversité des races de chiens, les erreurs de classification (faux positifs) permettent de comprendre les types de confusion qui persistent malgré l’optimisation du modèle. Les métriques choisies permettent également de comparer directement les performances du Visual Transformer avec celles du modèle de base VGG16 affiné.*

*##### Courbes de perte et de perte de validation*

*Les courbes de perte et de perte de validation offrent une vue d’ensemble de la convergence et de la stabilité du modèle. Une courbe de perte en diminution constante suggère que le modèle apprend correctement les caractéristiques du dataset, tandis qu’une perte de validation stable ou légèrement fluctuante reflète une généralisation adéquate. Pour le Visual Transformer, cette analyse permet de vérifier que le modèle ne sur-apprend pas les détails du dataset (overfitting), en maintenant une perte de validation stable sur les images inédites.*

*Courbes de précision et de précision de validation*

*Les courbes de précision et de précision de validation sont des indicateurs directs de la capacité du modèle à bien classer les images. La comparaison de la précision entre les jeux d'entraînement et de validation permet de juger de la généralisation du modèle. Pour le Visual Transformer, qui repose sur une approche différente de celle du CNN, l’analyse de ces courbes permet d’identifier des éventuelles faiblesses dans la capacité du modèle à différencier des races proches.*

*En résumé, l’analyse des faux positifs, des courbes de perte, et des courbes de précision fournit une interprétation exhaustive des performances du Visual Transformer. En remplaçant l’analyse de la feature importance par cette approche axée sur les métriques, nous obtenons une évaluation qualitative et quantitative de la capacité du modèle à classer précisément et de manière robuste les différentes races de chiens du Stanford Dogs dataset.*

**Les limites et les améliorations possibles**

*Limites du Vision Transformer*

*Besoins en données : ViT fonctionne mieux sur de grands volumes de données, limitant sa généralisation avec des datasets plus modestes comme Stanford Dogs.*

*Complexité : Le modèle exige plus de puissance de calcul que les CNN, ce qui peut poser des défis pour le déploiement.*

*Interprétation complexe : Bien que les cartes d'attention montrent des zones d'intérêt, l'interprétation des décisions du modèle reste difficile.*

*Améliorations envisageables*

*Augmentation de données : Enrichir la capacité de classification en ajoutant plus de races de chiens à classifier, étant donnée les très bonnes performances obtenues avec Visual Trasnformer pour 4 races de chien.*

*Approche hybride CNN-ViT : Associer des couches CNN avec des couches Transformer pour combiner caractéristiques locales et globales.*

*Optimisation : Régler les hyperparamètres comme la taille des patches et le dropout pour un meilleur compromis performance-généralisation.*

*Exploration de Transformers plus légers : Tester des architectures comme les Swin Transformers pour réduire les besoins en données.*