Note méthodologique : preuve de concept

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.

Chaque image sera redimensionnée sur 300x300, il s'agit plus ou moins des dimensions moyennes des images et elles produisent de meilleurs résultats qu'avec les dimensions usuelles de 224x224.

J'avais remarqué dans le projet précédent que pour l'entraînement de modèles personnalisés ajouter des étapes de préparation des images (réduction du bruit, égalisation de l'histogramme des images, ...) n'influençait pas les résultats, je ne vais donc pas en utiliser ici non plus. Des méthodes d'augmentation d'images diverses légères seront par contre utilisées, elles sont mentionnées plus tard dans cette note.

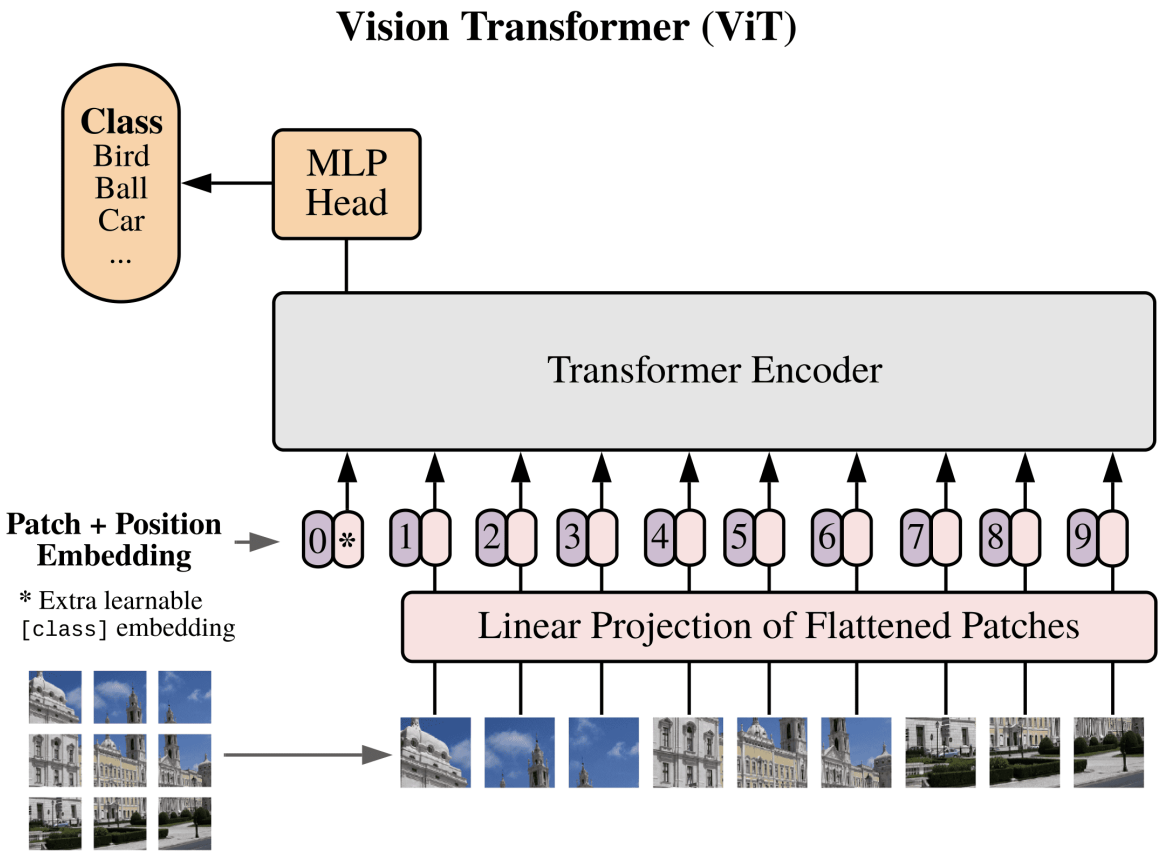
Les concepts de l'algorithme récent

Le mécanisme d'attention dans le contexte de la classification d'images permet au modèle de se concentrer sur les parties pertinentes d'une image pendant son traitement. Pour se faire ce mécanisme va attribuer un poids d'attention pour chaque partie/segment d'une image. Par exemple, dans notre cas de classification d'images de chien, le modèle pourrait attribuer un poids d'attention supérieur aux zones contenant des caractéristiques distinctives telles que les oreilles ou la forme générale du chien tout en accordant moins d'importance aux éléments de l'arrière-plan.

Ce mécanisme représente la base de l'architecture des t*ransformers*. Ces derniers ont révolutionné le traitement du langage naturel mais peuvent également être utilisés dans d'autres domaine, typiquement celui de la vision par ordinateur via l'architecture *Vision Transformers* (*ViT*).

Ces modèles *VIT* vont d'abord diviser une image en segment d'une taille fixe (typiquement 16x16 pixels). Leurs positions respectives vont également être encodées avec chaque segment afin de permettre aux réseaux de connaître la position exacte de chaque segment dans l'image initiale. Le mécanisme d'attention va ensuite être utilisé dans les couches de *Transformers* pour calculer l'attention entre les différentes parties de l'image.

Le diagramme suivant provenant du document de recherche "*An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale*" représente cette architecture VIT sous forme visuelle:



La modélisation

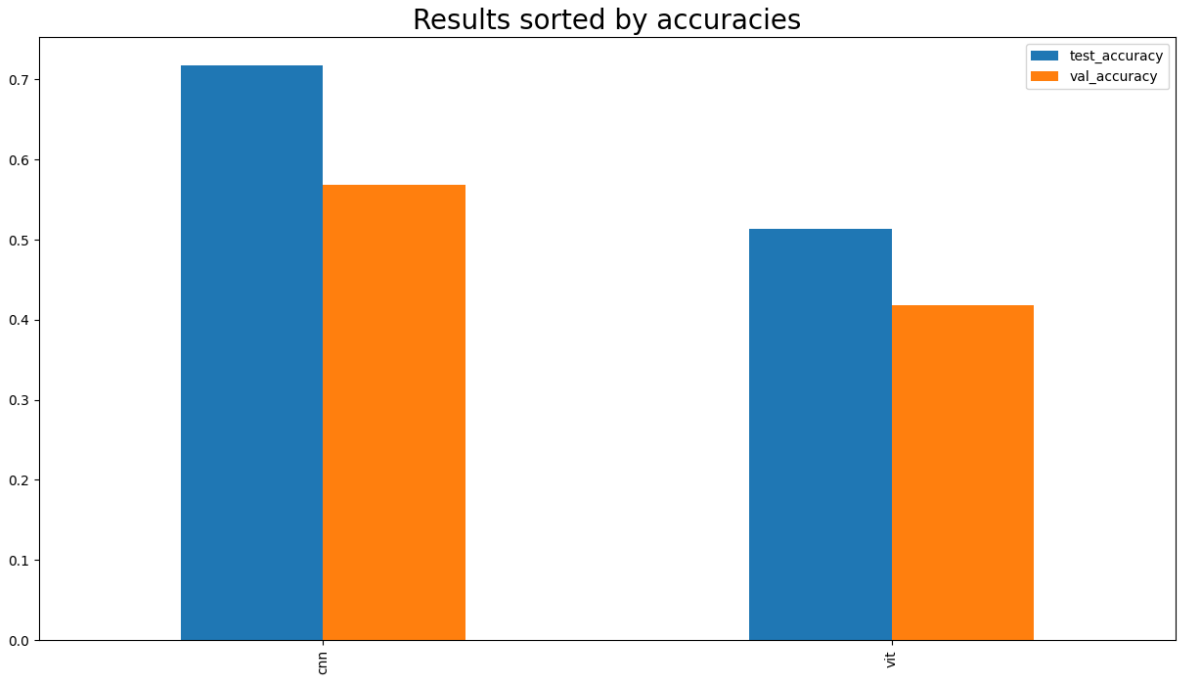
Le modèle VIT personnalisé comporte ces différentes couches :

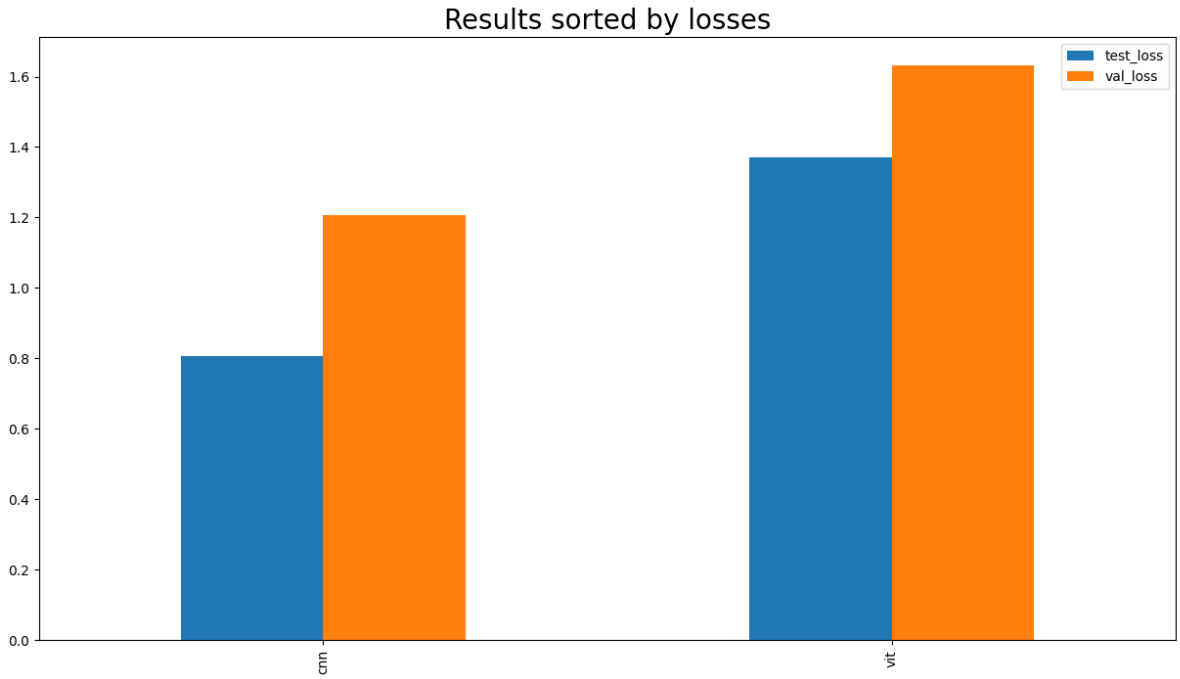
- Une première couche d'Input pour accepter les images de taille 300x300.
- Quelques couches d'augmentation des données. Les différentes augmentations sont :
  - Rotation aléatoire
  - Translation aléatoire
  - Flip horizontal aléatoire
  - Changement de contraste aléatoire
  - Zoom aléatoire
  - Changement d'éclairage aléatoire
- Les couches créant les segments des images, ces patches sont de taille 16x16.
- Une couche ajoutant les positions respectives de chaque segment.
- Les blocs de couches t*ransformers* qui utilisent le mécanisme d'attention. Ils comportent également des couches de *MLP*, permettant au modèle d'apprendre des transformations plus complexes via l'activation GELU.
- Des couches pour condenser les informations apprises des couches précédentes et les préparer pour la classification.
- Les couches de classifications qui vont déterminer la probabilité d'appartenir à chaque race de chien.

Les deux modèles vont être entraînés de la même façon. Ils utilisent l'optimiseur classique Adam avec un taux d'apprentissage initial de 0,001. J'utilise pendant l'entraînement le mécanisme de *ReduceLROnPlateau* pour que le taux d'apprentissage diminue quand le modèle a l'air de cesser d'apprendre et le mécanisme d'*EarlyStopping* avec une plus grande patience pour l'arrêter quand l'apprentissage a cessé d'être utile malgré un taux d'apprentissage réduit précédemment au lieu d'utiliser un nombre d'époques fixes. La taille du batch est de 16, un bon compromis pour mon ordinateur entre la vitesse d'exécution et les ressources de GPU utilisées.

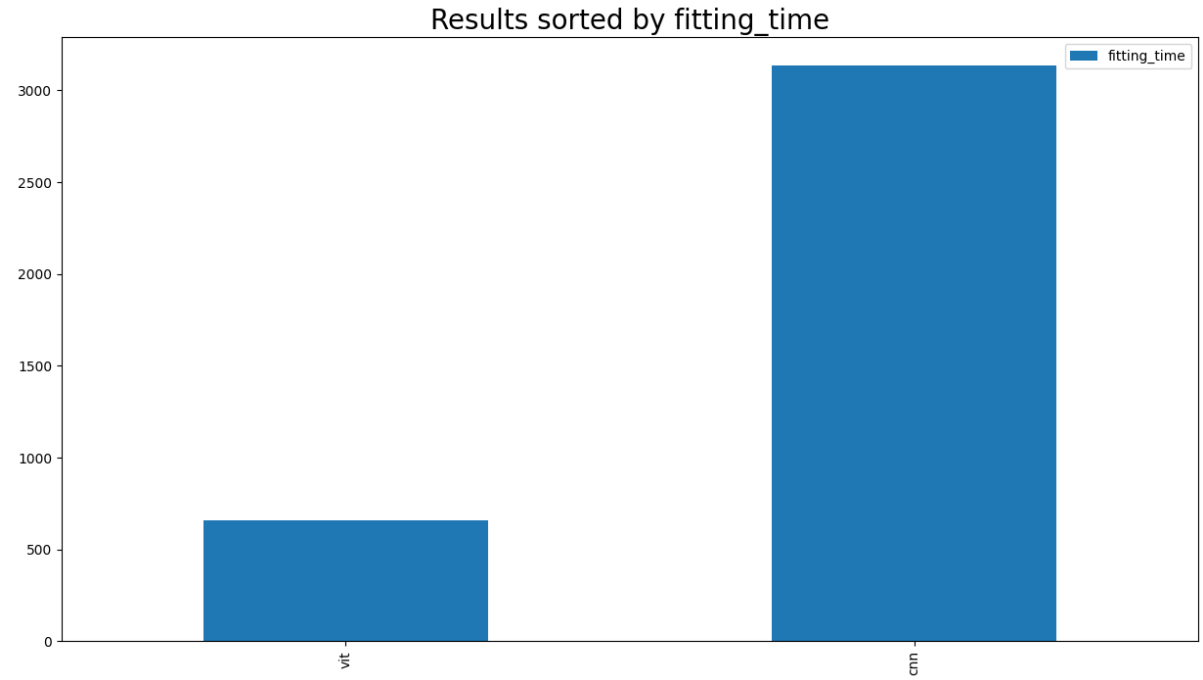
Une synthèse des résultats

Le modèle CNN a des résultats près de 50% meilleurs que le modèle VIT comme le montrent les tableaux suivants :





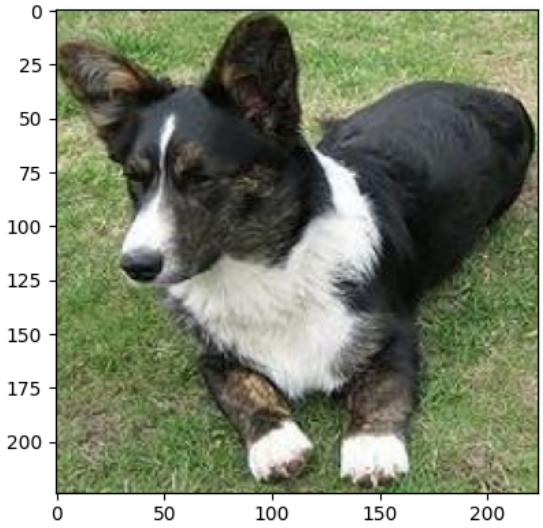
Cependant le modèle CNN a nécessité un temps d'entraînement considérablement plus grand :

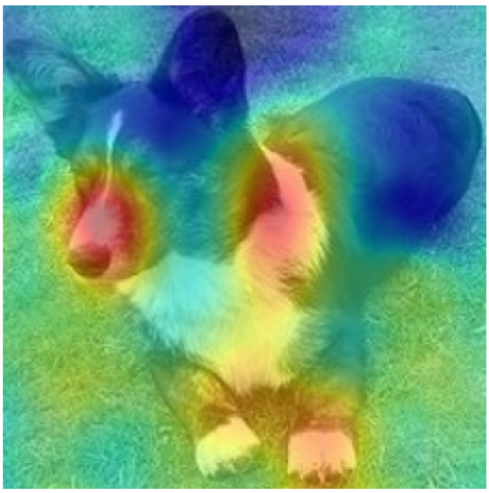


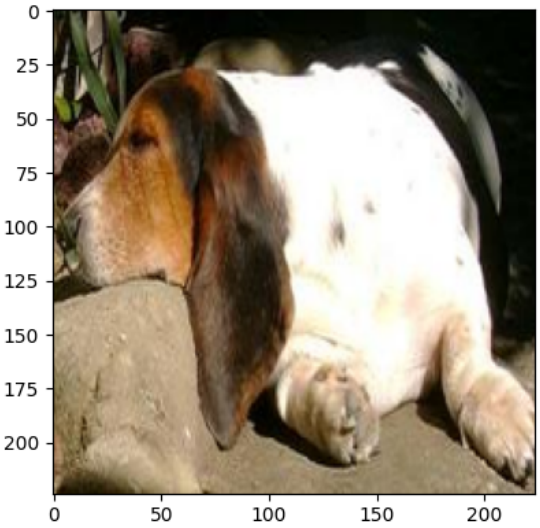
Les performances réduites du modèle VIT doivent provenir du fait que ces modèles nécessitent beaucoup de données pour être performants comparés au CNN, le dataset utilisé ici n'était clairement pas assez grand. Ce modèle cependant a eu besoin d'un temps d'entraînement bien plus court ce qui était prévu.

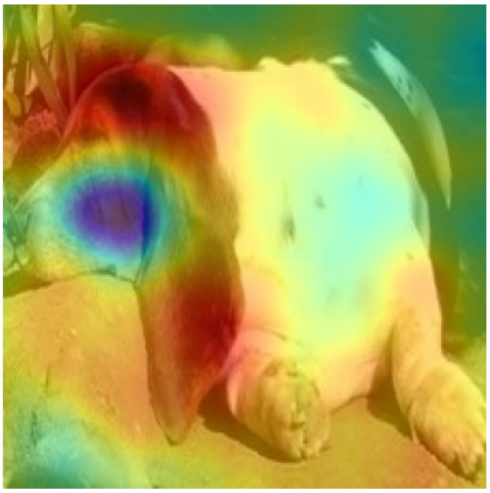
L’analyse de la feature importance globale et locale du nouveau modèle

La méthode RISE (Randomized Input Sampling for Explanation) permet de comprendre comment dans notre cas le modèle VIT a effectué sa décision de classification. Cette méthode va ici cacher aléatoirement certaines parties de l’image et observe à quel point le résultat est affecté pour déterminer quelles parties ont le plus d’importance. Les images suivantes représentent un exemple de cette méthode via une carte thermique de l’attention :









Les limites et les améliorations possibles

- Une des restrictions avec les modèles VIT est qu'ils nécessitent un nombre important de données pour être efficace. Ici il n'y a clairement pas beaucoup d'images par race de chien, cela explique les résultats plutôt pauvres obtenus avec le modèle VIT personnalisé. Utiliser un set de données plus important permettrait d'améliorer ses performances.

- Des modèles pré-entraînés offrent de meilleurs résultats vu qu'ils sont déjà entraînés sur des sets de données importants. Il suffit par la suite de les entraîner une nouvelle fois sur notre set de données via transfert d'apprentissage pour qu'ils offrent les meilleures performances possibles pour notre cas business. Utiliser un modèle pré-entraîné VIT permettrait d'améliorer les performances.

- Des architectures plus récentes ont montré des performances supérieures aux modèles VIT telles que les modèles DeiT (Data-efficient Image Transformers), BeiT (BERT pre-training of Image Transformers) ou ViTMAE (Masked Autoencoders).