

**Développez une preuve de concept**

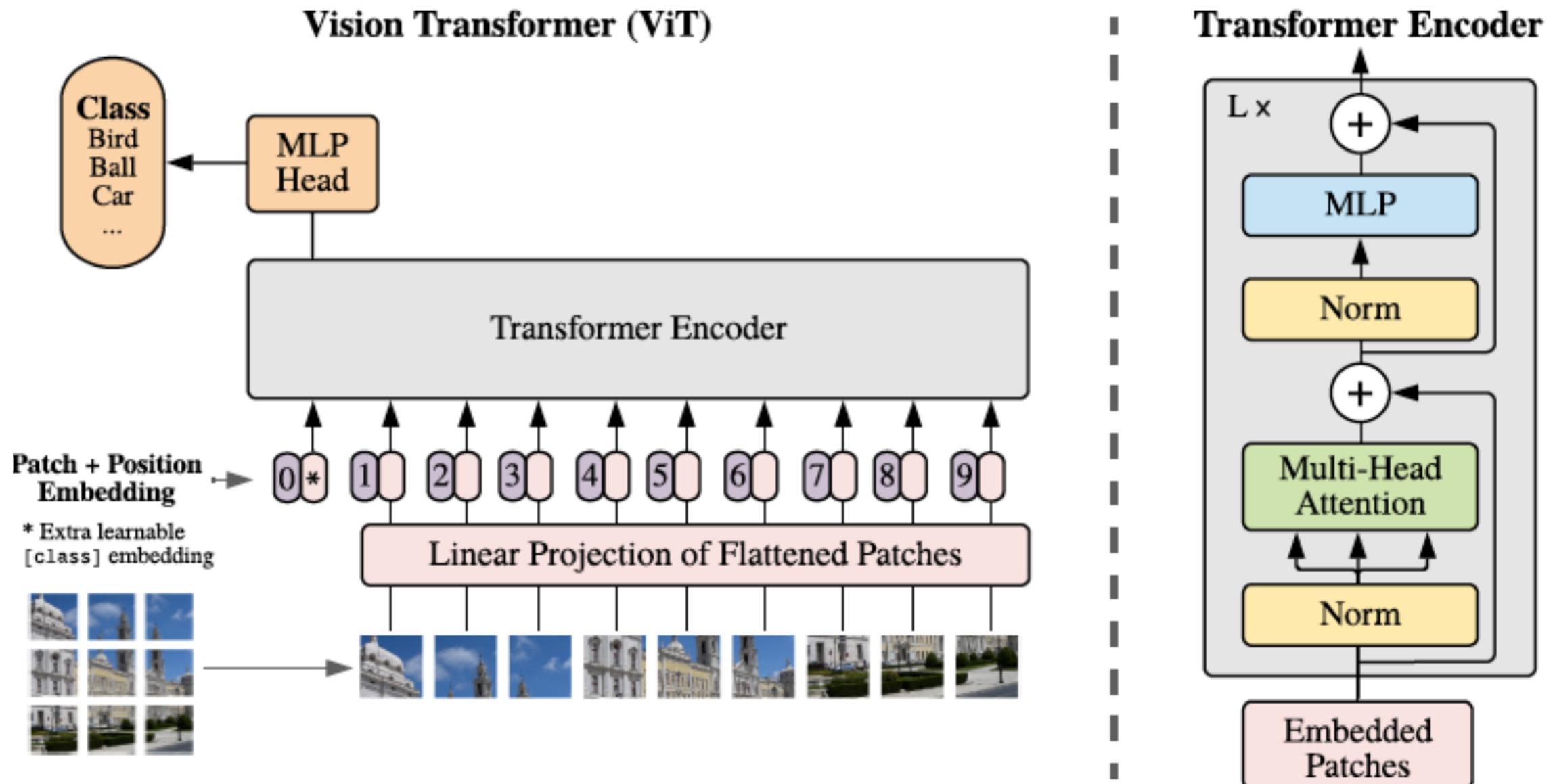
**Soutenance du projet n°7 : Parcours « Ingénieur Machine Learning »**

Luke Duthoit

## **Plan de la présentation :**

- Présentation de la thématique.
- Présentation du jeu de données choisi.
- Présentation des la *baseline* et de la nouvelle méthode.
- Comparaisons de performances.

# Un algorithme inspiré des transformateurs<sup>1</sup> pour faire de la classification d'images



1 : DOSOVITSKIY Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv: 2010.11929 (2020)

# Une méthode originale à la croisée entre des applications de *deep learning*.

- Fait de la classification d'image sans CNN.
- Mécanisme d'extraction des features inspiré de l'analyse de données textuelles... sans réseaux récurrents.
- Performances prometteuses suite à pré-entraînement sur très grandes bases de données (impossible à faire en local sur notre ordinateur).
- Fait l'objet d'intenses études de comparaison avec les CNN<sup>1</sup> dans de nombreux domaines.

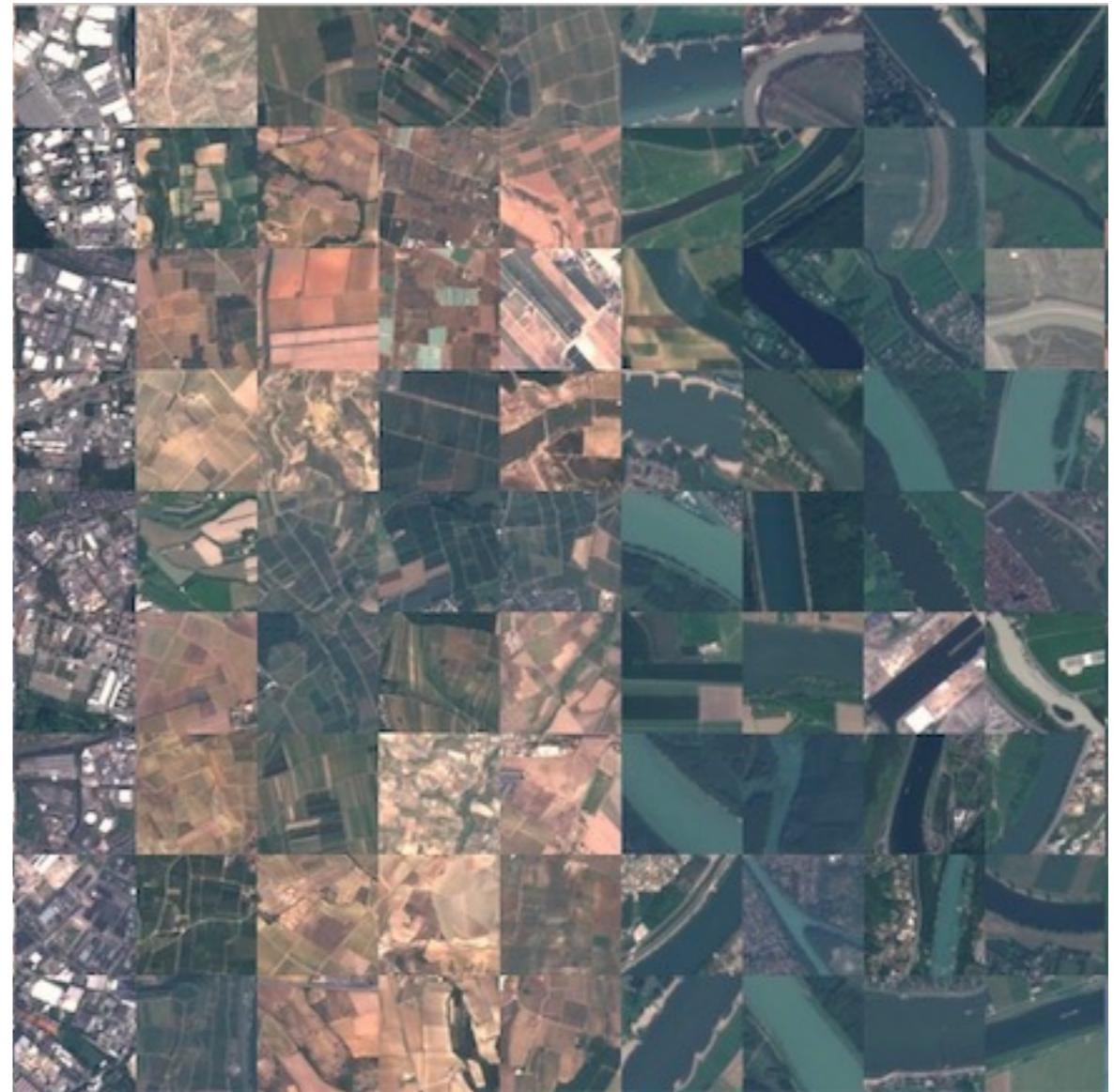
1 : M. RAGHU et al., Do Vision Transformers See Like Convolutional Neural Networks? arXiv:2108.08810 (2022)

## **Plan de la présentation :**

- Présentation de la thématique.
- Présentation du jeu de données choisi.
- Présentation des la *baseline* et de la nouvelle méthode.
- Comparaisons de performances.

# Un code inspiré d'un tutoriel...

- Tutoriel pour implémenter une architecture ViT dans un PB de classification d'image<sup>1</sup>.
- Librairies : *transformers*, *tensorflow.keras* et *datasets*.
- Tutoriel appliqué sur la base de données EuroSAT.



1 : Phil Schmid, <https://www.philschmid.de/image-classification-huggingface-transformers-keras>

# ... donc choix d'un nouveau *dataset* et d'une application différente

- Notre choix : re-utiliser Standard Dogs Datasets (projet n°6).
- Plus de 20k images (~EuroSat) de 120 races de chiens (»EuroSat).
-  *Baseline* naturelle : modèle du projet n°6.



# Pas la plus facile des base de données...

Nombreuses difficultés pour la reconnaissance de traits caractéristiques :

- ♦ Parfois plusieurs chiens par photo.



- ♦ Pas toujours de la même race.



- ♦ Plusieurs âges différents.

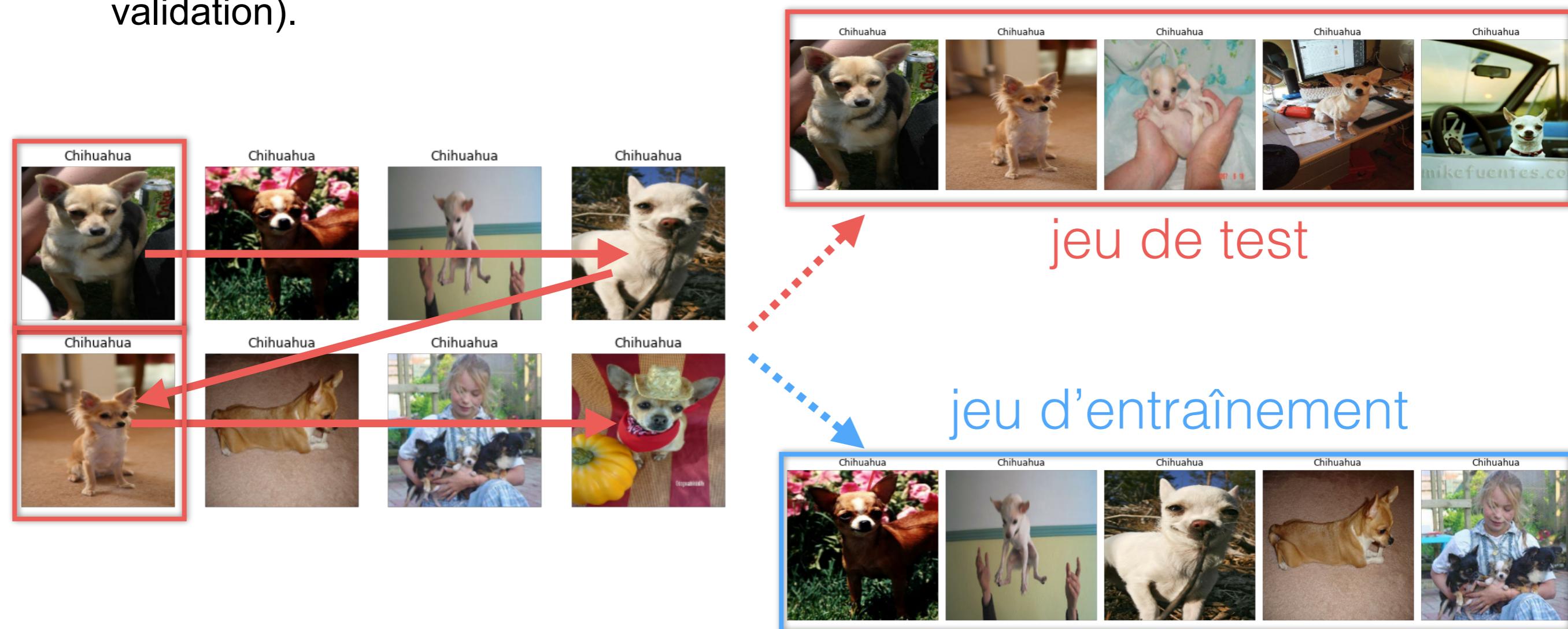


- ♦ Beaucoup de bruit autour du/des chien(s).



# Découpages en jeux d'entraînement / validation / test

- A cause de `image_dataset_from_directory`, impossible de respecter la séparation *train set / test set* indiquée en ressource de *Stanford Dogs Dataset*.
- Notre propre division : une image / quatre accordée au jeu de test (puis pour validation).



## ... d'autant qu'il faut réduire le nombre de classes

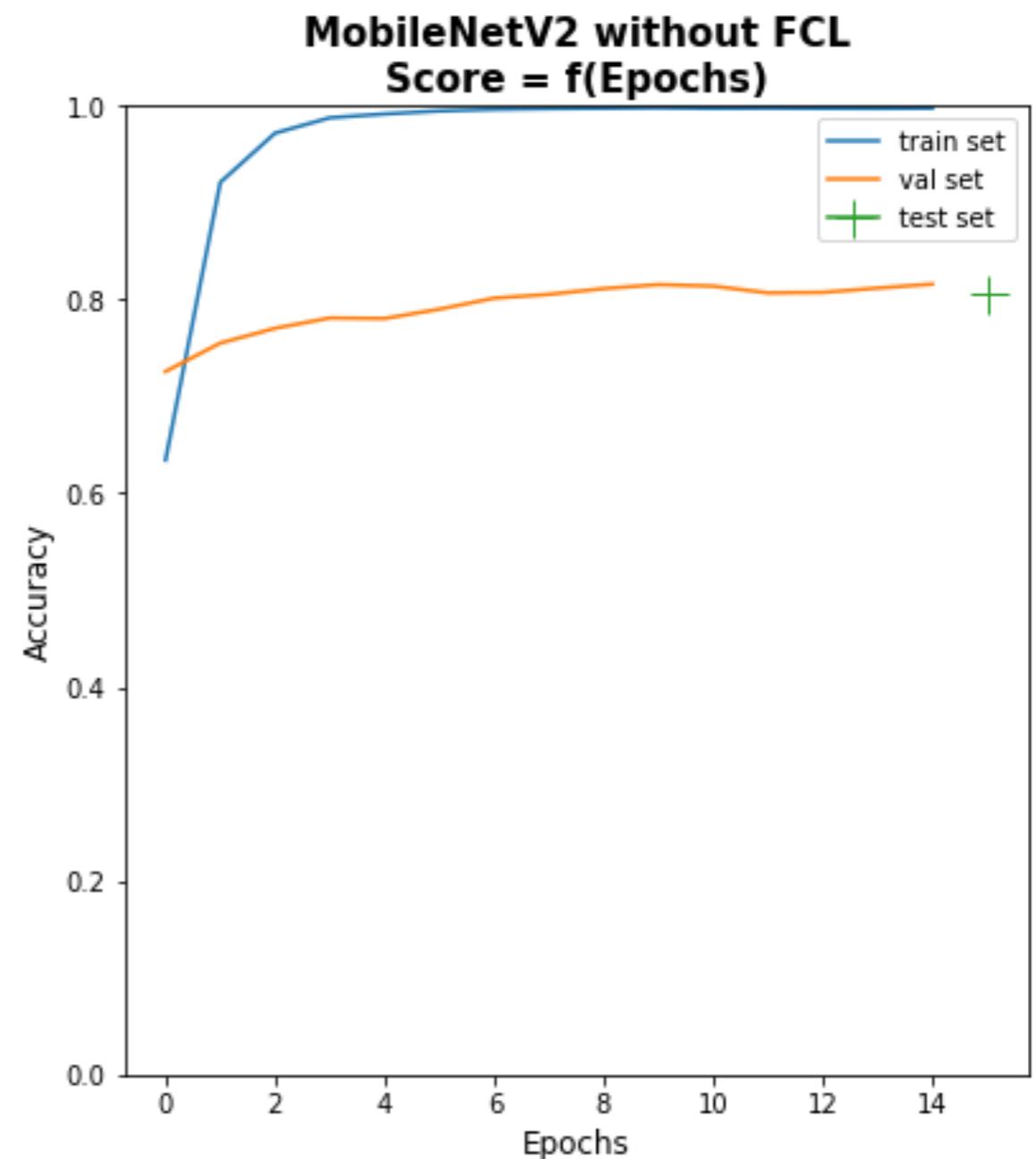
- ⚠️ Implémentation du code rend obligatoire d'avoir le même nombre d'image par classes.
- 🕒 Choix de ne retenir les 13 races contenant 150 images.
- Création « manuelle » d'une 2<sup>nde</sup> base de données à 28 races (15 races supplémentaires initialement à 151 ou 152 images/catégorie).

## **Plan de la présentation :**

- Présentation de la thématique.
- Présentation du jeu de données choisi.
- Présentation des la *baseline* et de la nouvelle méthode.
- Comparaisons de performances.

# La baseline : *transfert learning* sur MobileNetV2.

- Pré-traitement des images adapté.
- Extraction de *features* seulement :
  - pas de ré-entraînement de blocs convolutifs ;
  - suppression des FCL initiales ;
  - une nouvelles FCL à 120 sorties.
- Résultats honorables... sur 120 races  $\Rightarrow$  devra être ré-entraîné sur les nouveaux *datasets*.



## La nouvelle méthode : *transfert learning* sur **google/vit-base-patch16-224-in21k**

- Architecture de type ViT, pré-entraînée sur ImageNet-21k (10 fois + d'images et 20 fois + de classes qu'ImageNet).
- Images découpées en *patches* de 16x16.
- **ViTFeatureExtractor** pour le pré-traitement des images, et **TFViTForImageClassification** pour construire l'architecture.
- *Fine tuning* sur les nouvelles bases de données (près de 80 fois + de poids de connexions à re-apprendre que pour la *baseline*).
- Régulation des poids de connexion (*weight decay*).

## Points communs aux deux méthodes

- Hasard fixé (*seed constante*).
- Même *resizing* des images en entrée(224x224)
- Même nombre d'*epochs* (5 car faible nombre de races).
- Même regroupement des images par batch de taille 32.

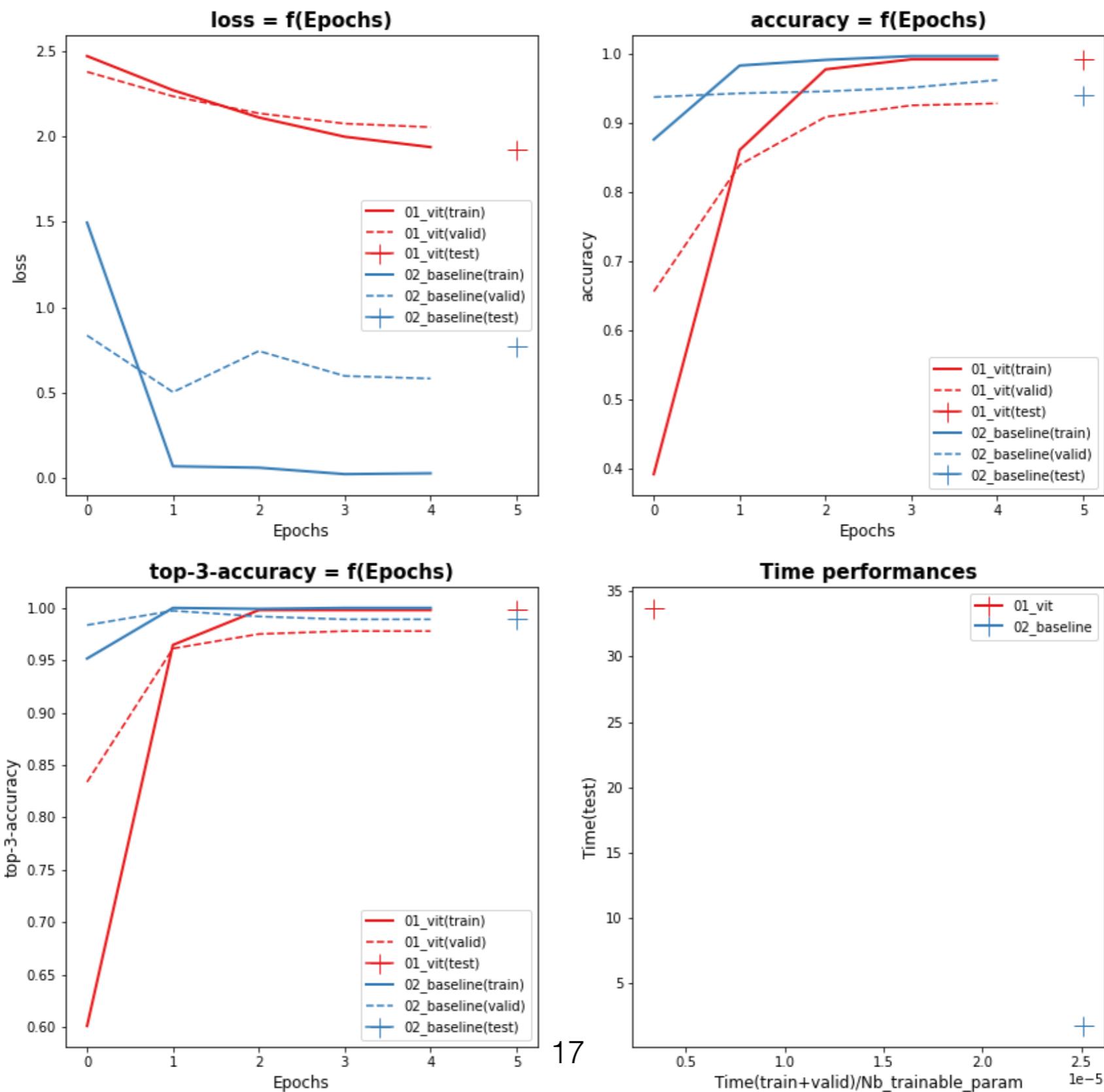
## **Plan de la présentation :**

- Présentation de la thématique.
- Présentation du jeu de données choisi.
- Présentation des la *baseline* et de la nouvelle méthode.
- Comparaisons de performances.

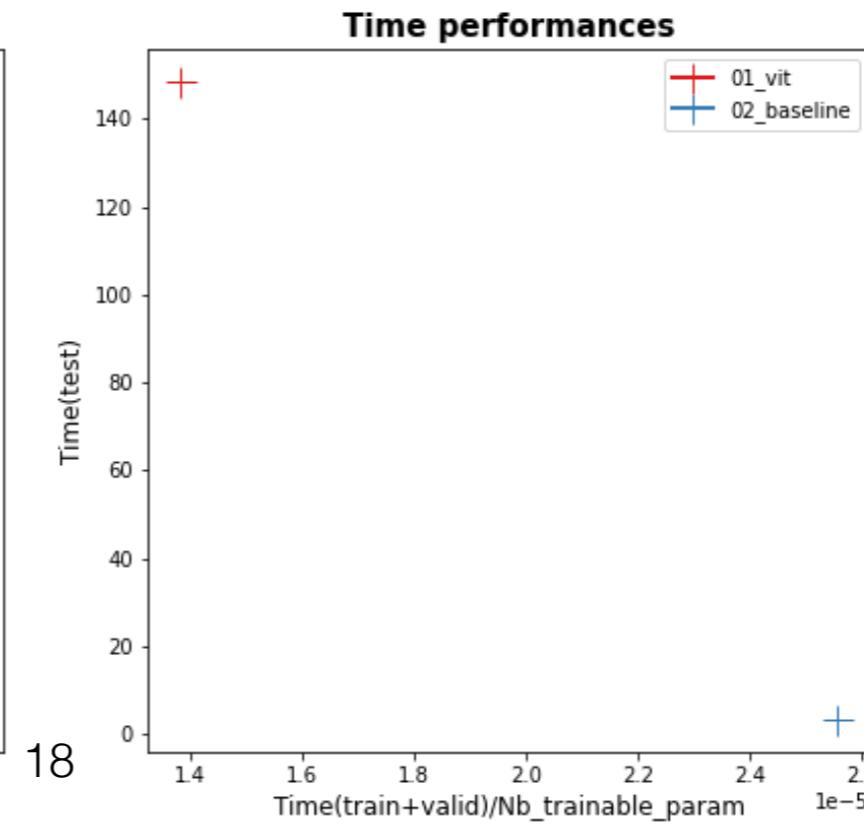
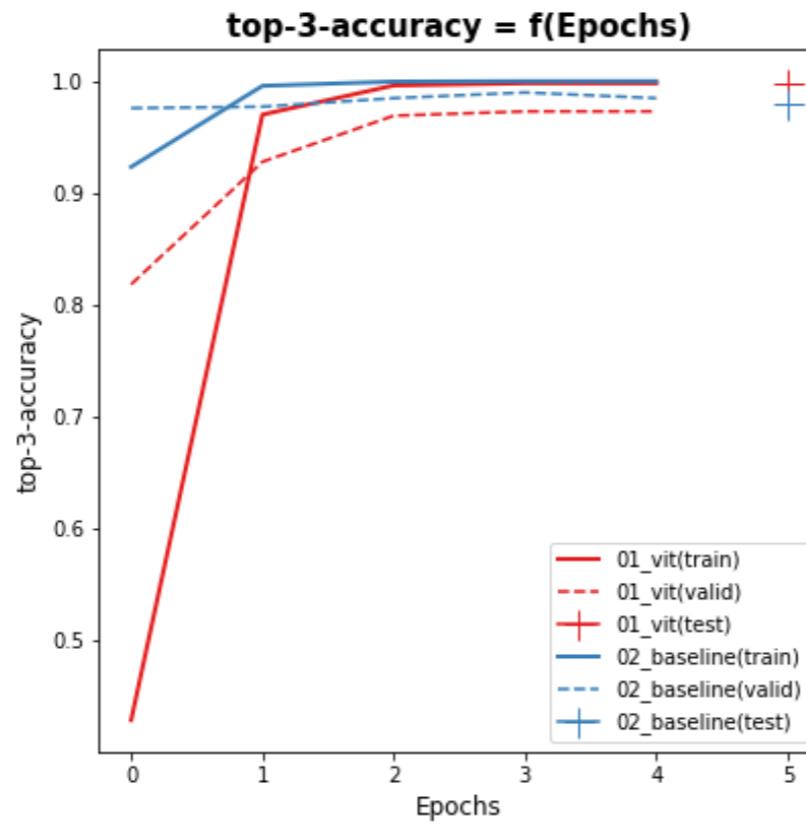
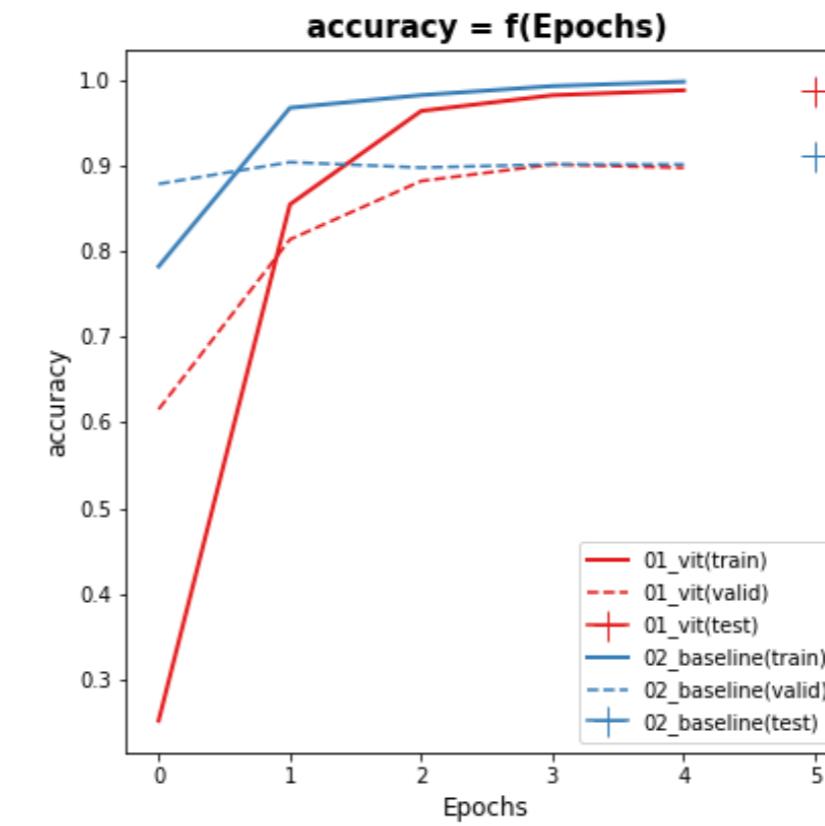
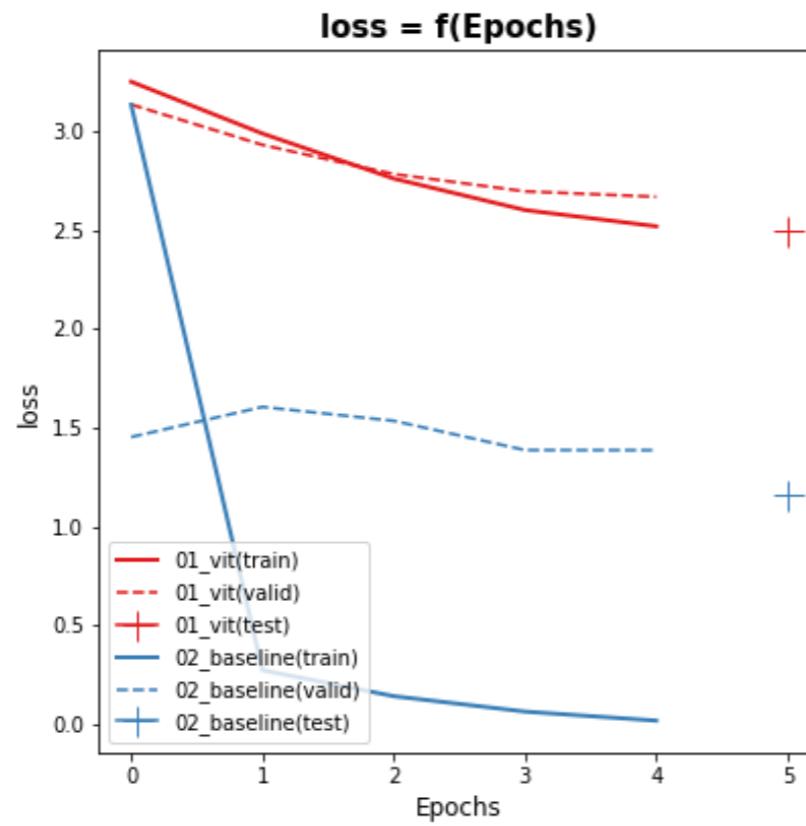
# Nos observables

- Fonction coût : entropie croisée.
- Métriques :
  - score de précision (*accuracy*);
  - top-3-précision ⇒ prend en compte quand l'étiquette fait partie des classes les plus probables
- Performances temporelles :
  - durée **absolue** de l'évaluation sur jeu de test
  - durée **relative** de l'entraînement/validation (*cf* différence de nombre de poids à apprendre...)

# Des performances très comparables sur 13 races...

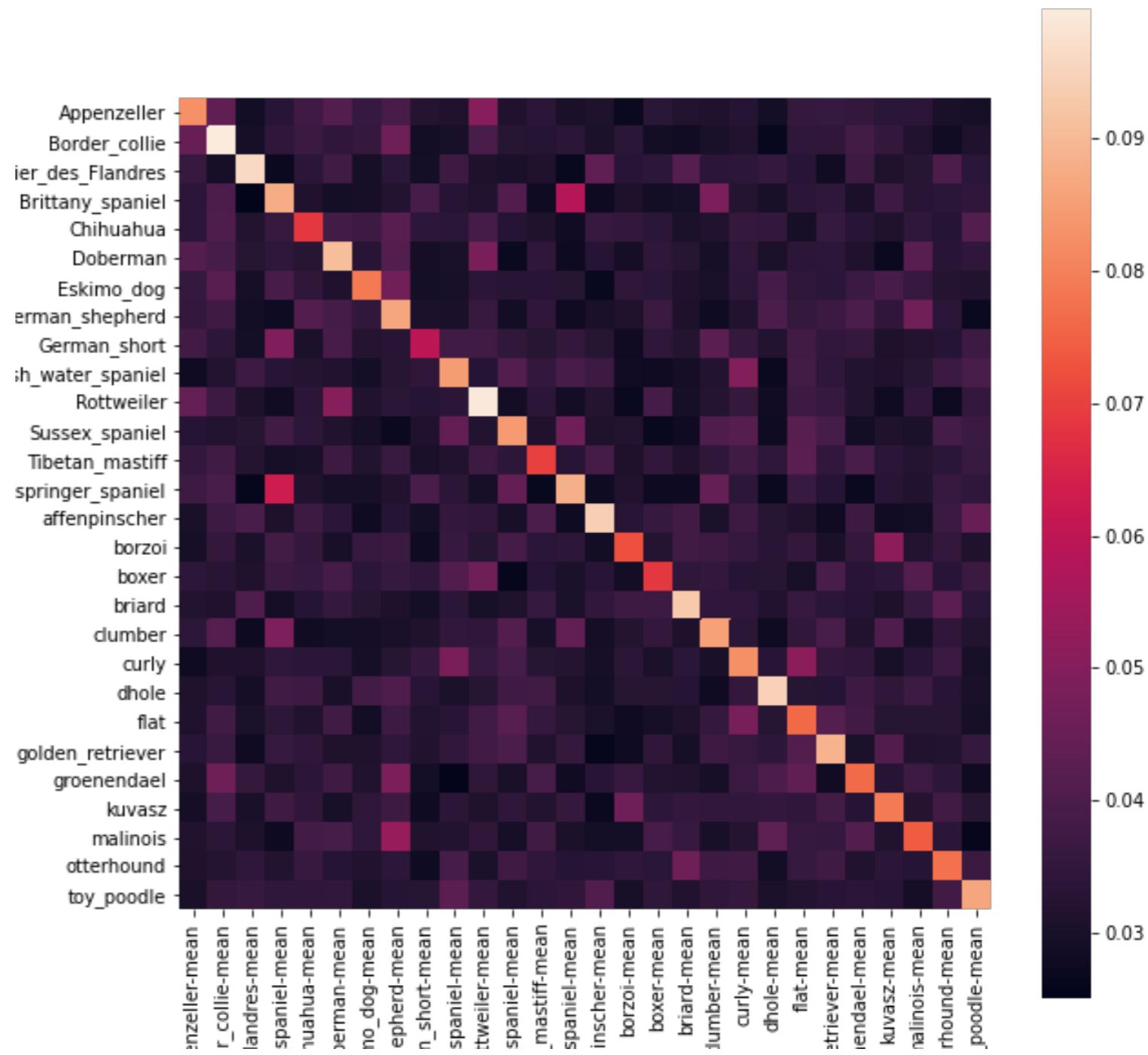


# ... reproduites sur 28 races

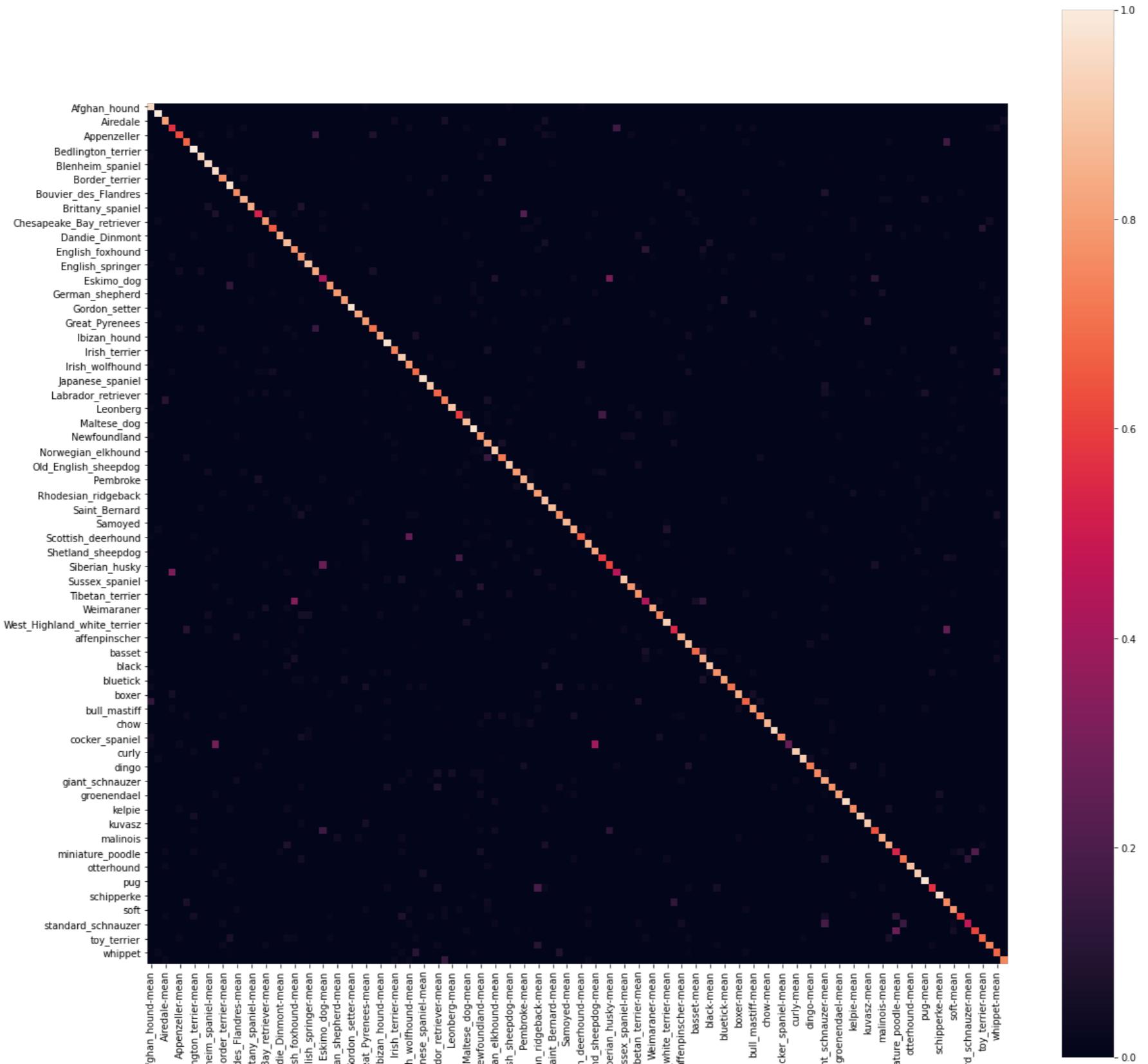


# Une matrice de confusion dominée par les termes diagonaux

- $M[i,j] = \langle P(\text{race}[j]) \rangle_{\text{étiquette}[i]}$
- Domination des termes diagonaux (cf bons scores)
- Mais faibles écarts avec coeff. hors diagonale  $\Rightarrow$  prédiction « seulement » suffisante



# Pour rappel : la matrice de la baseline... sur 120 races.



# Des confusions entre races qui souvent se ressemblent... ce qui explique et nuance les erreurs de l'algorithme

Race la moins bien prédite n° 1  
German\_short ( $<P>=0.06$ )



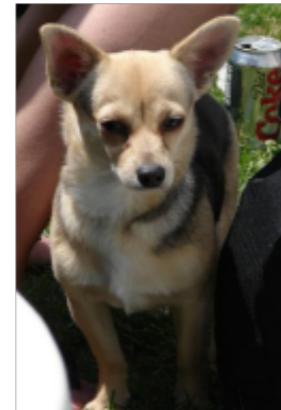
Souvent confondu avec  
Brittany\_spaniel ( $P=0.05$ )



Souvent confondu avec  
clumber ( $P=0.04$ )



Race la moins bien prédite n° 2  
Chihuahua ( $<P>=0.07$ )



Souvent confondu avec  
German\_shepherd ( $P=0.04$ )



Souvent confondu avec  
toy\_poodle ( $P=0.04$ )



Race la moins bien prédite n° 3  
boxer ( $<P>=0.07$ )



Souvent confondu avec  
Rottweiler ( $P=0.05$ )



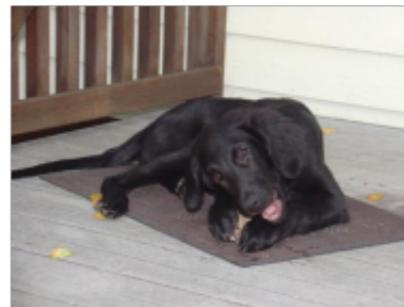
Souvent confondu avec  
malinois ( $P=0.04$ )



Race la moins bien prédite n° 4  
Tibetan\_mastiff ( $<P>=0.07$ )



Souvent confondu avec  
flat ( $P=0.04$ )



Souvent confondu avec  
groenendael ( $P=0.04$ )



Race la moins bien prédite n° 5  
borzoi ( $<P>=0.07$ )



Souvent confondu avec  
kuvasz ( $P=0.05$ )



Souvent confondu avec  
Sussex\_spaniel ( $P=0.04$ )



# Conclusions

- Nous avons implémenté un algorithme de classification d'images par transfert learning (fine tuning) à partir d'une architecture nouvelle (ViT).
- Comparé à notre baseline, la nouvelle méthode conduit à des performances de mêmes niveau (scores, et temps relatif d'entraînement)... sur 10% - 25% des classes.
- Probabilités de prédictions suffisantes mais relativement faibles.

# Perspectives

- Étendre le travail sur les 120 races possibles (puissance de calcul supplémentaire nécessaire).
- Avec du temps, chercher à implémenter le modèle entier avec keras<sup>1</sup>.
- Se renseigner sur les modèles « hybrides » entre ViT et CNN pour voir si les prédictions sont renforcées.

1 : [https://keras.io/examples/vision/image\\_classification\\_with\\_vision\\_transformer/](https://keras.io/examples/vision/image_classification_with_vision_transformer/)