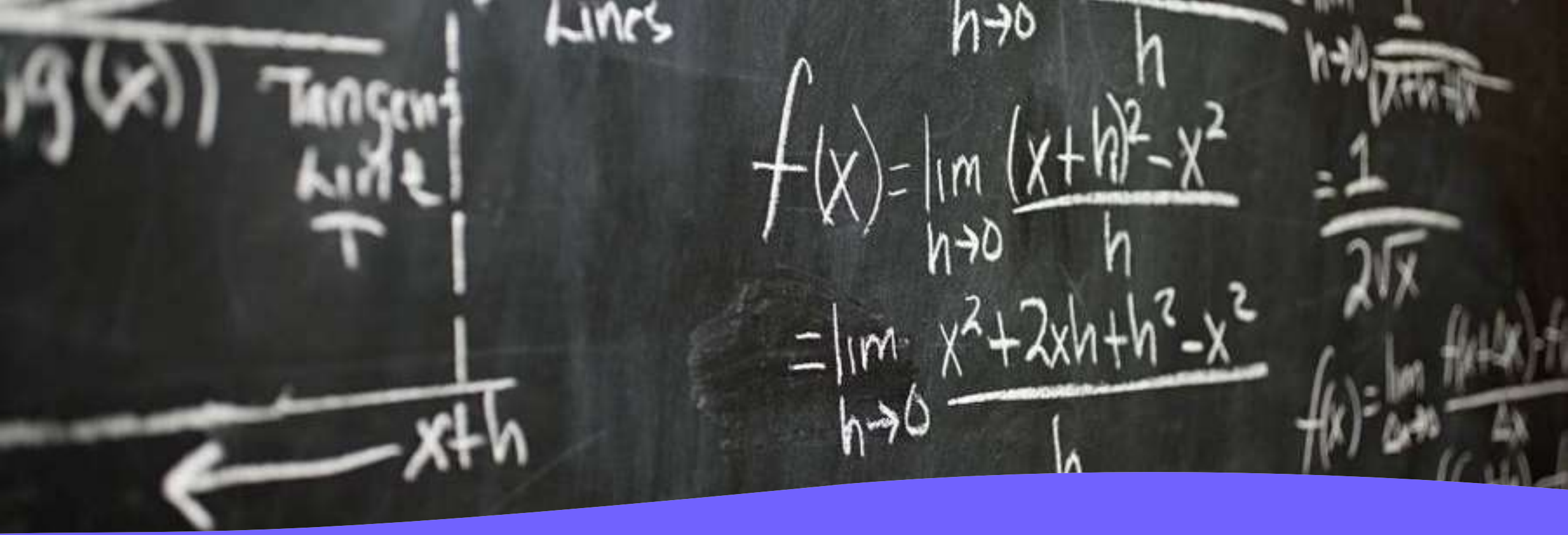
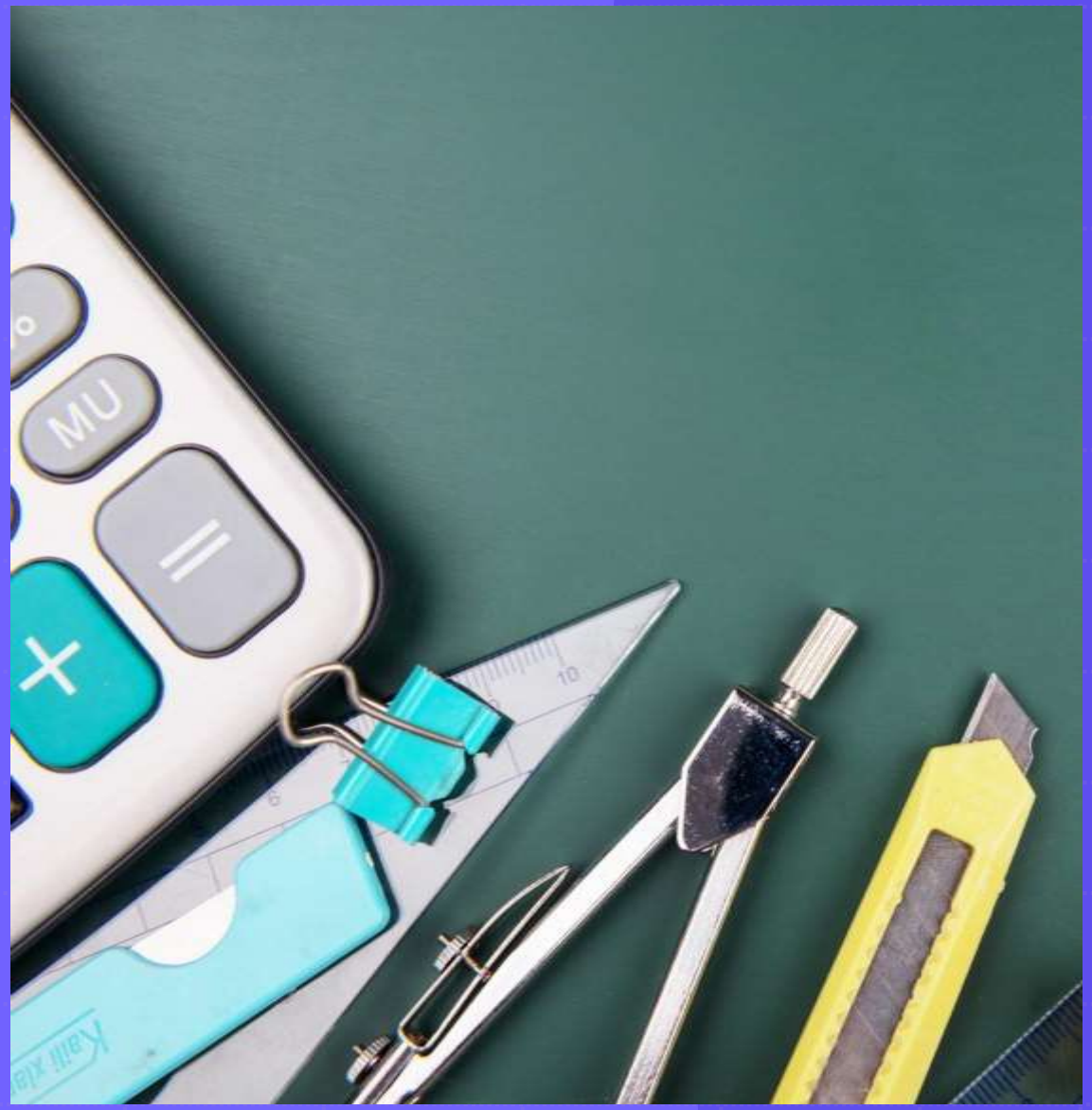


Développez une preuve de concept

BOURBON Vicente

Problématique

- Evolution très rapide du monde de la Data Science
- Besoin de pouvoir monter rapidement en compétences sur de nouvelles méthodes
- Mission: réaliser une preuve de concept d’une méthode récente (moins de 18 mois) pour la comparer à une méthode déjà en production



Sommaire

I. Cadre de travail

- Classification d'images avec CNN lors d'un précédent projet
- Recherche sur l'état de l'art et découverte d'une nouvelle méthode: ViT
- Objectif: mettre en œuvre un ViT et le comparer à un CNN (temps de calcul et Accuracy)
- Réutilisation du Stanford Dogs Dataset et du CNN Xception précédemment entraîné
- Modèle ViT-B/16 issu d'un article de recherche des équipes Google

Sources bibliographiques

- Article de recherche
- Code source
- Article de vulgarisation
- Tutoriel d'utilisation du modèle

II. Tour d’horizon de la classification d’images

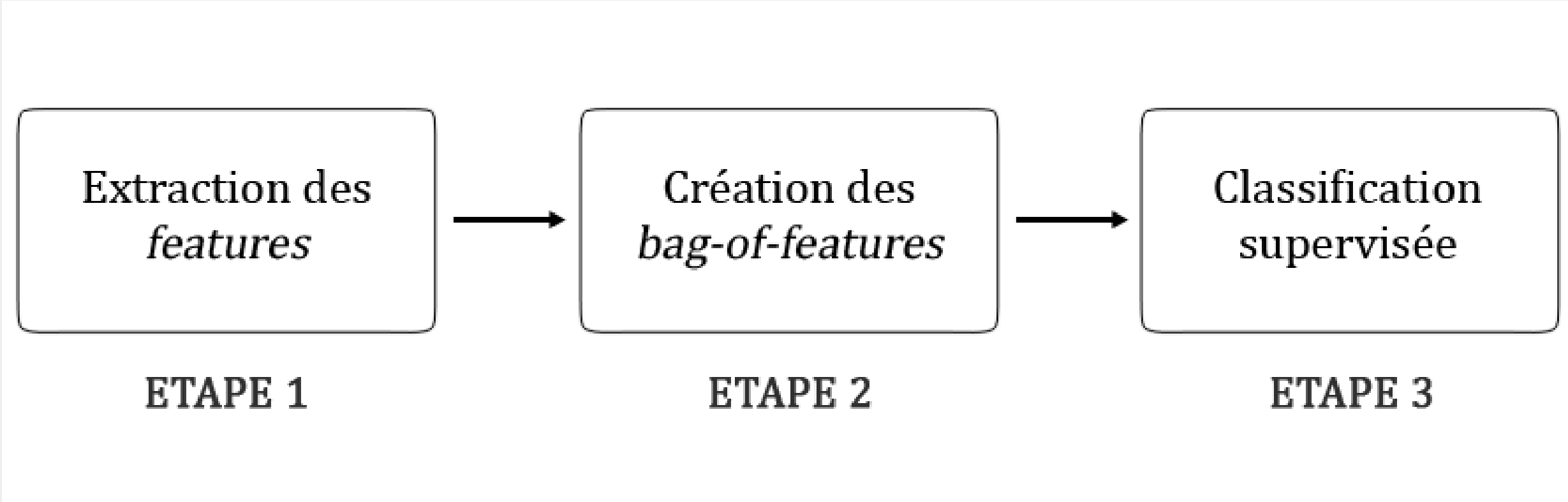
Présentation et état de l’art

- Rappel: image représentée par matrice de pixels
- Classification d’images: système d’assignation d’une catégorie à une image
- Classification supervisée et non supervisée
- Importance de la classification d’images avec l’explosion des données

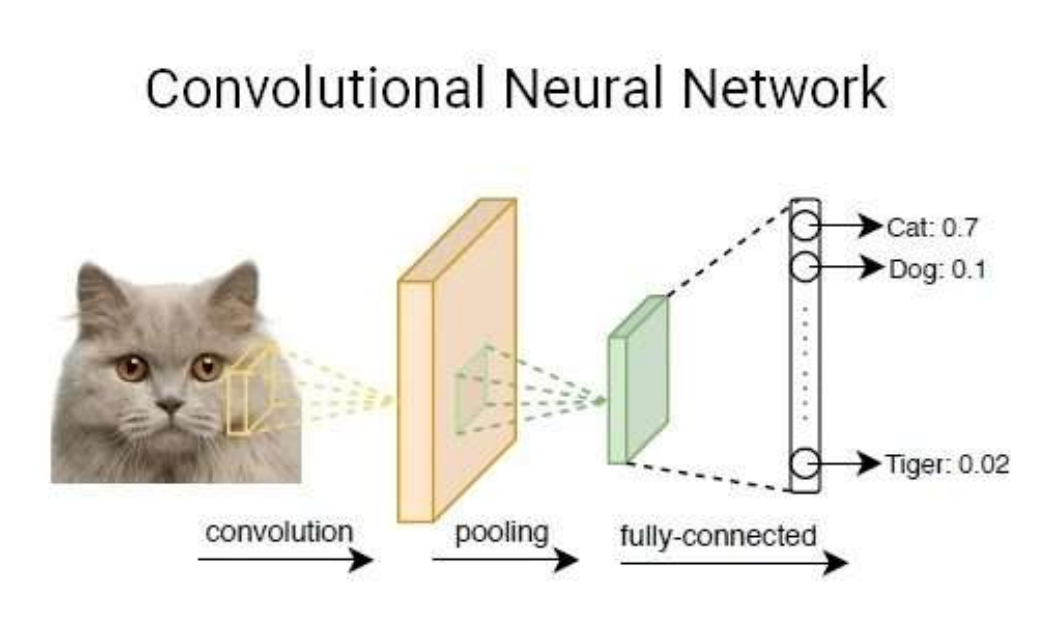
Evolution

- Premières méthodes début du siècle
- Révolution Deep Learning à partir de 2012

Fonctionnement général



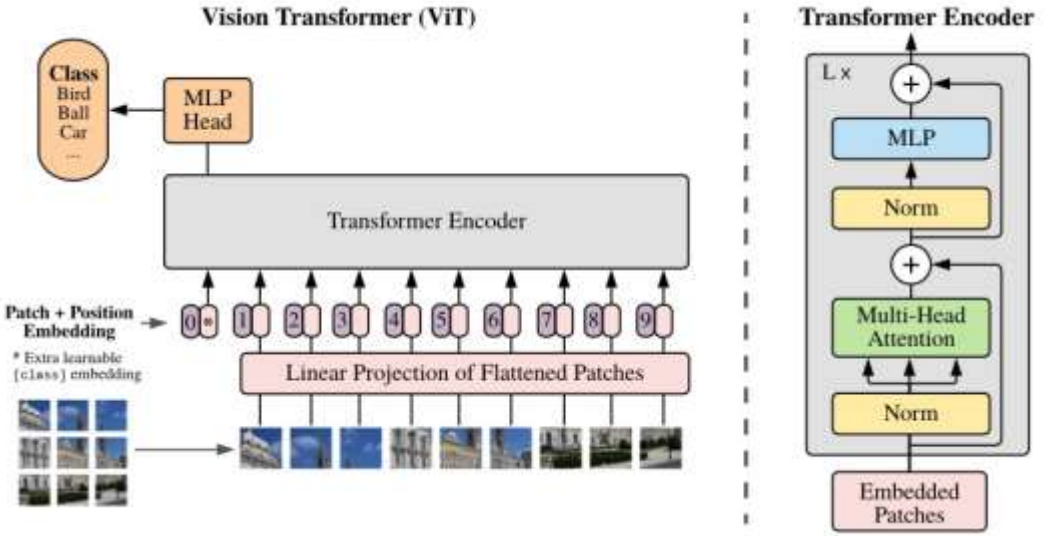
Fonctionnement d'un CNN



- Automatisation et optimisation des traitements
- Unité de base: perceptron
- Perceptrons regroupés en couches
- 4 types de couches:
  - Convolution: recherche de features
  - Pooling: réduction de la taille des images
  - Correction Relu: limiter le sur-apprentissage
  - Fully-connected: classification
- Xception: couches profondes séparables et connexions récurrentes

III. Vision Transformers

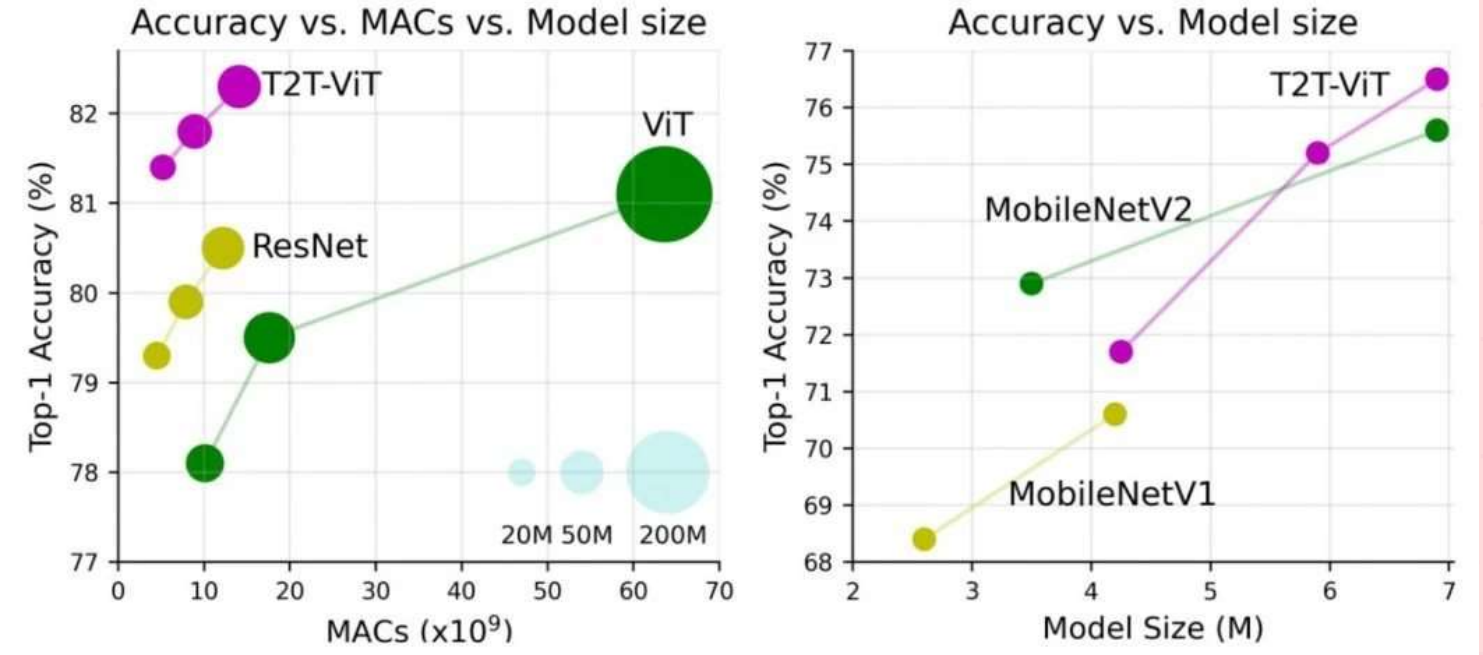
- ViT: modification d’un réseau d’encodage de Transformer
- Patching: division de l’image en patchs
- Aplatissement et transformation linéaire des patchs
- Numérotation des emplacements des patchs dans l’image originale
- Application d’un Transformer

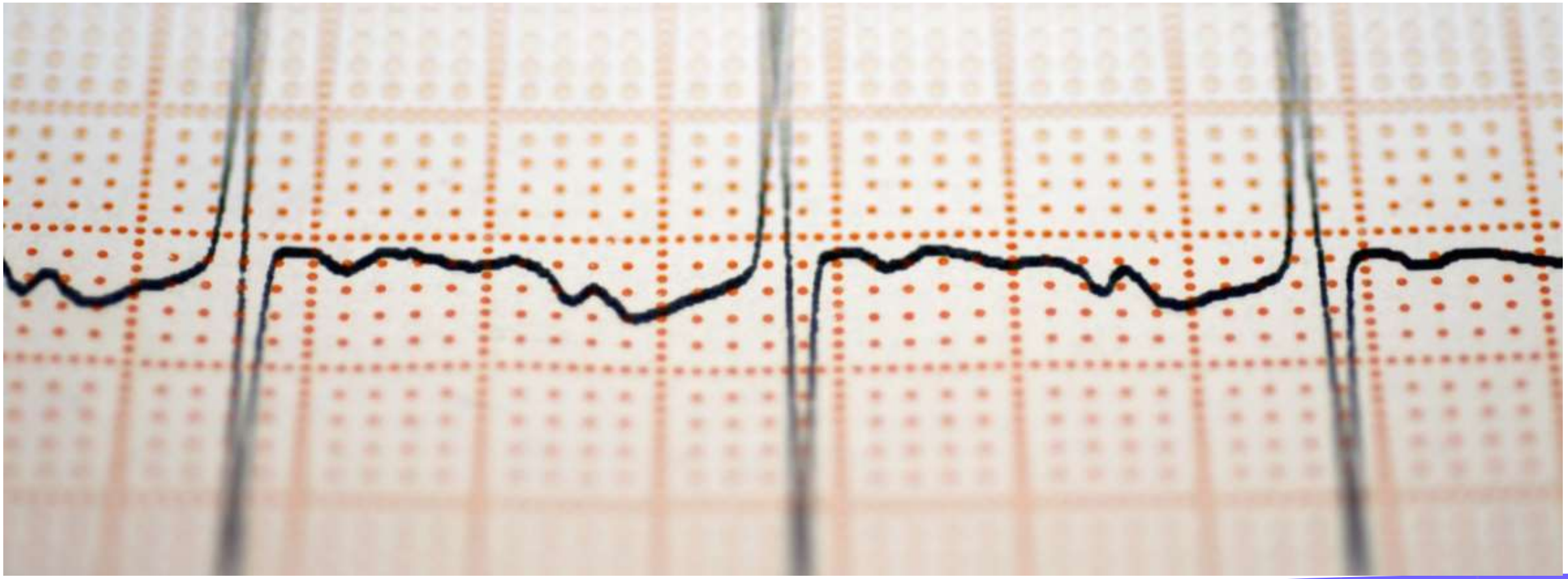


Architecture et fonctionnement

ViT vs CNN

- ViT n’a pas les mêmes capacités de généralisation qu’un CNN
- Moins bon résultats si entraîné sur un petit jeu de données
- Résultats équivalents, voir meilleurs, si entraînés sur un gros jeu de données
- Complexité de calcul plus faible





IV. Méthodes de classification

Entraînement des modèles

Xception

- Fine-tuning partiel: 10% des couches hautes de convolution réentraînées
- Optimisation des hyperparamètres
- Entraînement avec GPU

ViT-B/16

- Fine-tuning à partir d’un modèle disponible sur HuggingFace
- Hyperparamètres par défaut
- Entraînement avec GPU

Résultats

- Temps de calcul similaires
- Meilleure précision avec ViT
- ViT plus difficile d’utilisation mais de plus en plus populaire
- ViT intéressant pour transfer learning mais pas pour entraînement complet
- ViT nouvelle méthode en production

	Temps calcul (s)	Accuracy train	Accuracy test
Xception	4463	0.9878	0.7272
ViT	4566	0.9922	0.8243

V. Conclusion

- Nouvelles compétences:
  - Preuve de concept
  - Vision Transformers
- Principales difficultés:
  - Utilisation des modules et librairies HuggingFace
- Pistes d’amélioration:
  - Optimisation des hyperparamètres de ViT
  - Ajout de Data Augmentation