

Rapport de synthèse - IA pour l'image (Computer Vision)

Groupe : Dehmani Manar - Ab Der Halden Cyril - Blois Axel - Siarri Julie

Objectif : comparer plusieurs approches de classification d'images et retenir, pour chaque base, la solution offrant le meilleur compromis performance, généralisation et complexité de calcul, en s'appuyant sur les notions vues en cours (apprentissage supervisé, descripteurs, BoVW, SVM, CNN).

Organisation et répartition du travail

Le projet a été mené de manière collaborative par l'ensemble du groupe. Une première phase de travail commune a permis de mettre en place un pipeline partagé, notamment pour les réseaux de neurones convolutifs, afin d'harmoniser le chargement des données, les prétraitements, les métriques d'évaluation et la structure des notebooks.

Dans un second temps, le travail a été réparti en deux binômes afin de couvrir l'ensemble des approches étudiées :

- Le binôme Julie Siarri et Manar Dehmani a principalement travaillé sur les réseaux de neurones convolutifs et le Transfer Learning. Ce binôme a réalisé l'entraînement des modèles sur les différentes bases, l'étude de l'impact de la profondeur des réseaux, la mise en place des mécanismes d'optimisation (early stopping, sauvegarde des meilleurs poids, suivi des ressources) et l'analyse des courbes de performance.
- Le binôme Axel Blois et Cyril Ab Der Halden a principalement travaillé sur l'approche ORB combinée à Bag of Visual Words et SVM. Ils ont pris en charge l'extraction des descripteurs ORB, la construction des vocabulaires visuels, l'entraînement des classificateurs SVM, ainsi que l'analyse des matrices de confusion et des performances obtenues.

Les résultats ont ensuite été mutualisés et analysés collectivement afin de comparer les différentes approches et de déterminer, pour chaque base, la solution la plus adaptée.

0) Livrables attendus (rappel consigne)

Conformément à la consigne :

1. Trois notebooks (un par base) correspondant à la solution retenue, structurés et commentés, contenant des cellules Markdown explicatives et des cellules de code exécutées avec résultats affichés.
2. Trois fichiers HTML exportés à partir des notebooks. Les résultats ont également été fournis en PDF.

3. Un rapport de synthèse (ce document) présentant la description des bases, la justification des solutions retenues, les performances et la complexité, les raisons de rejet des autres méthodes, ainsi que l'environnement d'exécution.
-

1) Contexte et rappel théorique (cours)

La problématique traitée est une classification supervisée : apprendre une fonction de décision associant une image à une classe à partir d'exemples annotés. Cette formulation correspond au cadre vu en cours sur l'apprentissage supervisé et la classification.

Deux grandes familles d'approches ont été étudiées.

1.1 Approches par descripteurs (vision classique)

Le cours distingue la description d'images par descripteurs numériques de bas niveau (couleur, texture, points d'intérêt), en particulier les descripteurs locaux, puis leur agrégation via une représentation Bag of Visual Words et leur classification par des méthodes telles que les SVM.

ORB est un descripteur local rapide basé sur FAST et BRIEF.

BoVW consiste à construire un vocabulaire visuel par clustering (KMeans), puis à représenter chaque image par un histogramme de mots visuels.

Les SVM sont des classificateurs à vaste marge utilisés pour séparer les classes à partir de ces représentations.

1.2 Approches par features apprises (Deep Learning)

Les réseaux de neurones convolutifs apprennent automatiquement des représentations hiérarchiques à partir des données brutes. L'augmentation de la profondeur du réseau permet d'augmenter la capacité de représentation, au prix d'un risque accru de surapprentissage.

Le Transfer Learning s'inscrit dans cette logique en réutilisant un réseau pré-entraîné (par exemple sur ImageNet) comme extracteur de caractéristiques, afin d'améliorer la généralisation lorsque les données disponibles sont limitées.

2) Description des bases d'images

2.1 Dogs vs Cats

Objectif : classer une image en chien ou chat.

Difficultés : variations de pose, d'éclairage, de fond et de cadrage ; risque de surapprentissage si le modèle exploite des biais de contexte.

2.2 MNIST

Objectif : reconnaître des chiffres manuscrits de 0 à 9.

Difficultés : faibles, dataset très structuré ; adapté pour tester l'efficacité d'un modèle sans complexité de scène.

2.3 Intel Image Classification

Objectif : classer des scènes naturelles en six catégories.

Difficultés : forte diversité intra-classe, confusions sémantiques entre certaines classes, variations importantes de textures et de compositions.

3) Protocole expérimental et critères de sélection

3.1 Métriques

Accuracy, F1-score macro et matrices de confusion.

3.2 Critères de choix

Le meilleur modèle est celui qui :

1. présente une tendance de généralisation avec une validation qui baisse ou se stabilise durablement,
2. atteint un minimum de validation exploitable,
3. montre un alignement cohérent entre entraînement et validation,
4. conserve une complexité computationnelle acceptable (temps et ressources).

3.3 Optimisations

Afin de garantir un apprentissage stable, reproductible et cohérent avec les objectifs de généralisation, plusieurs mécanismes d'optimisation ont été mis en place lors des expérimentations.

L'early stopping a été utilisé pour interrompre automatiquement l'entraînement lorsque la performance sur l'ensemble de validation cessait de s'améliorer. Ce mécanisme permet d'éviter le surapprentissage, en particulier lorsque la loss d'entraînement continue de diminuer alors que la loss de validation se stabilise ou augmente. Le critère de décision repose sur la tendance globale de la validation et non sur des fluctuations ponctuelles. Lorsque l'early stopping est déclenché, les poids correspondant au meilleur epoch observé sur la validation sont restaurés.

La sauvegarde des meilleurs poids (checkpoint) a été systématiquement activée. Elle permet de conserver le modèle associé à la meilleure performance de validation, indépendamment de la durée totale de l'entraînement. Ce choix garantit la reproductibilité des résultats, facilite l'analyse a posteriori des performances et permet de comparer équitablement différentes architectures ou configurations.

Un suivi des ressources a également été mis en place afin d'analyser la complexité computationnelle des modèles. Le temps d'entraînement par epoch, l'utilisation du CPU, de la mémoire RAM et, lorsque disponible, de la mémoire GPU ont été mesurés. Cette instrumentation permet de ne pas limiter l'analyse à la performance de classification seule, mais de lier explicitement les gains de performance au coût de calcul associé.

Ces optimisations s'inscrivent dans une démarche de sélection raisonnée des modèles, en accord avec les notions vues en cours, où le choix final repose sur un compromis entre performance, capacité de généralisation et complexité de calcul.

4) Solutions étudiées

Les solutions étudiées ont été choisies de manière à couvrir différentes familles de méthodes en vision par ordinateur, conformément aux notions abordées en cours, tout en s'appuyant sur des pratiques couramment rencontrées dans la littérature et la documentation technique.

4.1 CNN from scratch

Les réseaux de neurones convolutifs ont été retenus comme approche principale, car ils constituent la méthode de référence actuelle pour les tâches de classification d'images. Cette approche est largement présentée en cours dans le cadre du deep learning appliqué à la vision par ordinateur, notamment pour sa capacité à apprendre automatiquement des représentations hiérarchiques à partir des données brutes.

Deux architectures ont été testées : un CNN à deux convolutions et un CNN à trois convolutions. Ce choix permet d'étudier l'impact de la profondeur du réseau sur la performance et la capacité de généralisation du modèle. L'objectif est de comparer un modèle plus simple, potentiellement suffisant pour des bases peu complexes, à un modèle plus profond, capable de capturer des structures visuelles plus riches, tout en évaluant le risque de surapprentissage.

Ce type de comparaison est directement inspiré des principes vus en cours sur le compromis entre capacité du modèle, biais et variance, et est également largement documenté dans les ressources en ligne et les tutoriels de référence.

4.2 ORB + BoVW + SVM

L'approche ORB combinée à une représentation Bag of Visual Words et à un classifieur SVM a été choisie comme méthode de vision dite classique, reposant sur des descripteurs manuels plutôt que sur des caractéristiques apprises automatiquement.

Cette méthode est présentée en cours dans le cadre de l'indexation et de l'analyse d'images, notamment pour illustrer les pipelines traditionnels basés sur l'extraction de points d'intérêt, la quantification des descripteurs et la classification supervisée. ORB a été retenu en raison de sa rapidité, de sa robustesse aux rotations et de son caractère libre d'utilisation, contrairement à d'autres descripteurs plus anciens.

Le modèle BoVW associé à un SVM constitue une chaîne classique largement utilisée avant l'essor du deep learning. Ce choix permet de disposer d'une base de comparaison pédagogique afin d'évaluer les apports réels des approches deep learning par rapport aux méthodes traditionnelles, tant en termes de performance que de complexité.

4.3 Transfer Learning

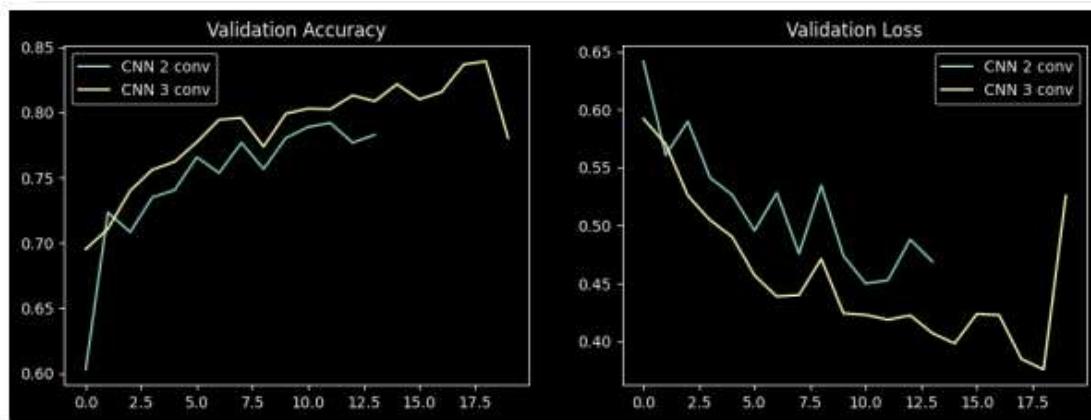
Le Transfer Learning a été étudié afin d'exploiter des représentations visuelles déjà apprises sur de grandes bases de données génériques, comme ImageNet. Cette approche est abordée en cours comme une solution efficace lorsque la quantité de données disponibles est limitée ou lorsque l'on souhaite accélérer la convergence des modèles.

Le réseau MobileNetV2 a été choisi car il est largement utilisé dans la littérature et dans la documentation officielle comme un compromis entre performance et coût computationnel. Il est fréquemment recommandé dans les ressources en ligne pour des tâches de classification d'images sur des bases de taille moyenne.

L'objectif de cette approche est d'évaluer si la réutilisation de caractéristiques pré-entraînées permet d'améliorer la généralisation par rapport à un CNN entraîné from scratch, et de comparer ce gain potentiel au coût computationnel associé.

5) Résultats et solution retenue par base (performances + complexité)

5.1 Dogs vs Cats - solution retenue : CNN à 3 convolutions



CNN 2 conv

- val_accuracy max environ 79.2 %
- val_loss min environ 0.45
- EarlyStopping : arrêt à l'epoch 14, restauration du meilleur epoch autour de 11
- Temps environ 11 à 12 secondes par epoch, RAM environ 15 à 16 Go, VRAM inférieure à 1 Go

CNN 3 conv

- val_accuracy max environ 83.9 %
- val_loss min environ 0.38
- Convergence plus tardive mais plus robuste, avec un meilleur minimum de validation
- Temps et VRAM comparables au CNN a 2 convolutions

Transfer Learning

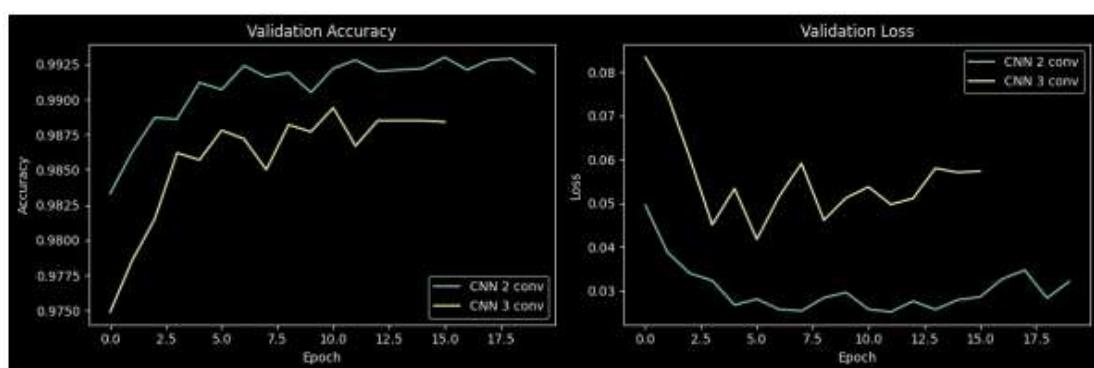
- val_accuracy max environ 76.6 %
- val_loss entre environ 0.49 et 0.56
- Temps environ 58 secondes par epoch, cout de calcul plus eleve pour une performance inferieure

ORB + BoVW + SVM

- Accuracy test environ 69.3 %
- F1 macro environ 0.69
- Couts : extraction ORB, construction du vocabulaire KMeans, representation BoVW et classification SVM, avec plusieurs etapes CPU et une consommation RAM comprise entre environ 5 et 6.7 Go

Justification du choix : Le CNN a 3 convolutions est retenu car il presente la meilleure performance de validation ainsi que le minimum de val_loss le plus bas, tout en conservant un cout de calcul similaire au CNN a 2 convolutions. Le Transfer Learning et l'approche ORB sont donc domines, soit par des performances inferieures, soit par un cout computationnel plus eleve.

5.2 MNIST - solution retenue : CNN à 2 convolutions



CNN 2 conv

- val_accuracy max environ 99.30 %
- val_loss min environ 0.025
- Temps environ 11 a 12 secondes par epoch, VRAM environ 0.8 Go

CNN 3 conv

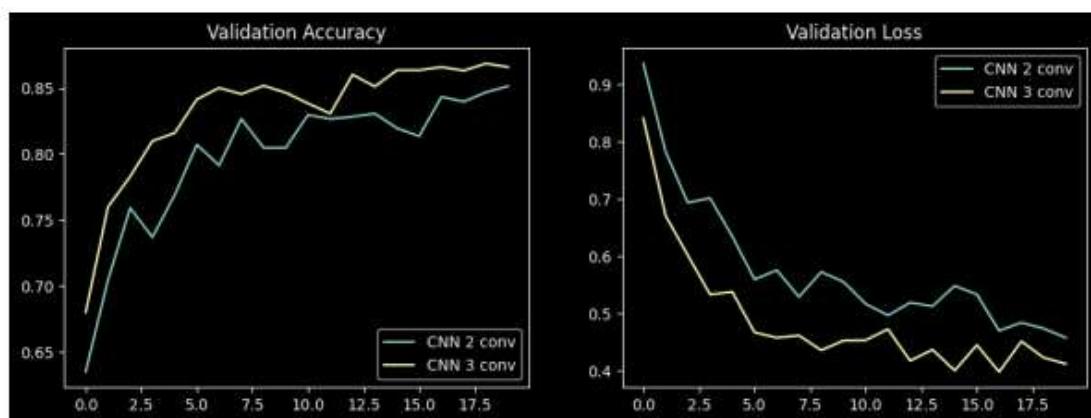
- val_accuracy max environ 98.94 %
- val_loss min environ 0.042
- EarlyStopping avec restauration du meilleur epoch
- Complexité plus élevée sans gain de performance

ORB + BoVW + SVM

- Accuracy environ 63.2 %
- F1 macro environ 0.62
- Méthode peu adaptée à MNIST en raison du faible nombre de points d'intérêt informatifs

Justification du choix : Le dataset MNIST étant simple et très structure, le CNN a 2 convolutions est suffisant pour atteindre une performance optimale. L'augmentation de la profondeur n'apporte pas de gain, tandis que l'approche ORB est largement moins performante.

5.3 Intel Image Classification - solution retenue : CNN à 3 convolutions



CNN 2 conv

- val_accuracy max environ 85.15 %
- val_loss min environ 0.4574
- Temps environ 49 secondes par epoch, VRAM inférieure à 1 Go

CNN 3 conv

- val_accuracy max environ 86.86 %
- val_loss min environ 0.4229
- Convergence plus lente mais meilleure capacité de généralisation
- Temps environ 51 à 54 secondes par epoch, VRAM inférieure à 1 Go

Transfer Learning

- val_accuracy max environ 80.0 %
- val_loss environ 0.55

- Temps environ 95 à 100 secondes par epoch, coût temporel plus élevé sans gain de performance

ORB + BoVW + SVM

- Accuracy test environ 48.8 %
- F1 macro environ 0.48
- Confusions importantes entre classes proches visuellement

Justification du choix : Le CNN a 3 convolutions offre le meilleur compromis entre performance de validation, capacité de généralisation et coût de calcul. Le Transfer Learning est plus lent sans amélioration notable, et l'approche ORB montre des limites importantes sur ce dataset multi-classes.

6) Difficultés rencontrées et solutions apportées

6.1 Difficultés liées aux CNN

Les courbes de validation se sont révélées parfois bruitées, en particulier sur Dogs vs Cats et Intel. Certaines architectures montraient une amélioration rapide sur l'entraînement sans amélioration durable de la validation.

Le choix de la profondeur du réseau a nécessité plusieurs expérimentations. Un réseau trop peu profond présentait une capacité insuffisante, tandis qu'un réseau plus profond augmentait le risque de surapprentissage.

Pour résoudre ces difficultés, l'analyse s'est appuyée sur la tendance globale des courbes plutôt que sur des valeurs ponctuelles. Une convergence tardive mais stable, avec un minimum de validation exploitable et un alignement cohérent entre entraînement et validation, a été privilégiée.

6.2 Difficultés liées à l'approche ORB + BoVW + SVM

L'approche ORB a présenté un coût computationnel important, notamment lors de l'extraction des descripteurs et de la construction du vocabulaire visuel par KMeans, exécutés sur CPU.

De plus, la combinaison de descripteurs locaux et de la représentation BoVW s'est révélée limitée pour capturer l'information sémantique globale. Sur Intel, cela se traduit par de fortes confusions entre classes proches. Sur MNIST, le faible nombre de points d'intérêt informatifs rend cette approche peu adaptée.

Ces limitations expliquent les performances inférieures observées et justifient l'exclusion de cette méthode des solutions finales.

7) Illustration des résultats dans les notebooks et HTML

Dans chaque notebook correspondant à une solution retenue :

- courbes train / validation,
- métriques finales,
- matrice de confusion,
- mesures de complexité,
- justification du modèle et de l'epoch retenu.

Les résultats détaillés des modèles non retenus restent disponibles dans les notebooks et fichiers HTML fournis.

8) Environnement d'exécution et dépendances

Python 3.10 ou supérieur, environnement Windows ou Linux, CPU multi cœur, GPU optionnel.

Librairies principales : numpy, matplotlib, opencv-python, scikit-learn, tensorflow/keras, psutil, gputil.

9) Conclusion générale

Les résultats montrent que les CNN offrent la meilleure capacité de généralisation sur l'ensemble des bases étudiées, à condition d'adapter la profondeur du réseau à la complexité des données.

Le Transfer Learning constitue une alternative intéressante mais n'apporte pas systématiquement un gain.

L'approche ORB + BoVW + SVM, bien que cohérente sur le plan théorique, atteint rapidement ses limites en performance et en coût.

Solutions retenues :

- Dogs vs Cats : CNN à 3 convolutions
- MNIST : CNN à 2 convolutions
- Intel : CNN à 3 convolutions

In []: