

**Classification d’images avec MobileViTv2**

**Note méthodologique & preuve de concept**

Sommaire

1. Dataset retenu

*(À compléter)*

1. Les concepts de l’algorithme récent
   1. Comparaisons

Bien sûr ! Voici la traduction en français, avec une explication ajoutée sur l’**auto-attention (self-attention)** :

**Qu’est-ce qu’une Convolution ?**

Une **convolution** est une opération mathématique largement utilisée en apprentissage profond, notamment dans les réseaux de neurones convolutifs (**CNN**), pour extraire des caractéristiques à partir de données telles que des images. Ce processus consiste à :

* Appliquer une petite matrice appelée **filtre** ou **noyau** qui « glisse » sur les données d’entrée (par exemple une image).
* À chaque position, calculer un produit scalaire entre le filtre et la partie correspondante de l’entrée.
* Cette opération met en évidence certains motifs (comme les contours ou les textures) et condense l’information, produisant une **carte de caractéristiques** qui résume les aspects importants de l’entrée.
* Les convolutions sont efficaces pour les images car elles utilisent des poids partagés et une connectivité locale, ce qui réduit le nombre de paramètres et permet de capturer des hiérarchies spatiales.

**Qu’est-ce qu’un Transformer ?**

Un **transformer** est une architecture d’apprentissage profond conçue pour traiter des données séquentielles (comme du texte ou des patchs d’images) en capturant les relations entre les éléments, quelle que soit leur distance dans la séquence. Les transformers se caractérisent par :

* L’utilisation de **mécanismes d’auto-attention** (self-attention) pour attribuer une importance différente à chaque partie des données d’entrée.
* La capacité à traiter des séquences entières en parallèle, ce qui les rend plus rapides et plus évolutifs que les anciens modèles comme les RNN.
* Une structure encodeur-décodeur, même si certains modèles n’utilisent qu’une seule partie.
* Ils sont devenus essentiels en traitement du langage naturel et sont désormais largement utilisés en vision par ordinateur sous la forme de **Vision Transformers (ViTs)**.

**Qu’est-ce que l’Auto-attention (Self-attention) ?**

L’**auto-attention** est un mécanisme qui permet à un modèle de « regarder » toutes les parties d’une séquence pour chaque élément de cette séquence. Concrètement, pour chaque élément d’entrée (par exemple, un mot dans une phrase ou un patch d’image), le modèle calcule à quel point il doit prêter attention aux autres éléments pour mieux comprendre le contexte global. Cela permet de modéliser efficacement les relations à longue distance et d’extraire des informations globales, ce qui est difficile avec des convolutions classiques qui ne voient que localement.

**Pourquoi la combinaison des deux dans MobileViTv2 est-elle efficace ?**

**MobileViTv2** combine convolutions et transformers pour exploiter les avantages des deux :

* Les **convolutions** sont excellentes pour capturer les motifs locaux et les hiérarchies spatiales dans les images, tout en étant très efficaces computationnellement pour les appareils mobiles.
* Les **transformers** (et en particulier l’auto-attention) sont très performants pour modéliser les dépendances à longue distance et le contexte global, ce qui manque souvent aux convolutions seules.

**MobileViTv2** est efficace car :

* Il utilise les convolutions pour l’extraction initiale des caractéristiques et le traitement local.
* Il intègre un mécanisme d’auto-attention allégé et séparable (une version plus efficace de l’attention des transformers) pour modéliser le contexte global.
* Il réduit la complexité computationnelle de l’attention classique, rendant le modèle plus rapide et adapté aux appareils à ressources limitées comme les smartphones.

Cette approche hybride permet à MobileViTv2 de surpasser les CNN purs et les premiers vision transformers en termes de précision et de rapidité, surtout sur du matériel mobile.

**Résumé : Convolution vs. Transformer dans MobileViTv2**

| **Aspect** | **Convolution** | **Transformer (Auto-attention)** | **Combiné dans MobileViTv2** |
| --- | --- | --- | --- |
| Ce qui est capturé | Motifs locaux | Dépendances globales | Les deux |
| Efficacité | Élevée (images) | Moins efficace (standard) | Optimisé pour le mobile |
| Nombre de paramètres | Faible | Plus élevé (attention standard) | Équilibré |
| Usage | Extraction locale | Modélisation du contexte global | Traitement image complet |
| Avantage MobileViTv2 | Rapidité, compacité | Représentations globales riches | Précision + rapidité |

La conception de MobileViTv2 lui permet d’offrir de hautes performances avec une faible latence et une taille de modèle réduite, ce qui le rend particulièrement adapté aux tâches de vision sur mobile.

**Comparaison de ResNet50 et MobileViTv2 pour le transfert d’apprentissage sur un petit jeu d’images (750 photos)**

**1. Fonctionnement des architectures**

**ResNet50**

* **Type** : Réseau de neurones convolutionnel profond (CNN).
* **Principe** : Utilise des couches convolutives et des « blocs résiduels » pour permettre un apprentissage efficace même dans des réseaux très profonds.
* **Points forts** :
  + Capture très bien les motifs locaux et les hiérarchies spatiales.
  + Très robuste grâce à la profondeur et aux connexions résiduelles.
* **Limite** :
  + Peut sur-apprendre (overfitting) facilement sur de petits jeux de données si la régularisation et l’augmentation de données ne sont pas bien utilisées[1](https://datascience.stackexchange.com/questions/47966/over-fitting-in-transfer-learning-with-small-dataset)[2](https://www.reddit.com/r/MLQuestions/comments/1j0tb9f/resnet50_cant_test_well_on_small_dataset_at_all/).

**MobileViTv2**

* **Type** : Architecture hybride combinant convolutions (pour l’extraction locale) et transformers (pour le contexte global).
* **Principe** : Mélange couches convolutives légères et blocs d’auto-attention pour capturer à la fois les détails locaux et les relations globales dans l’image.
* **Points forts** :
  + Très efficace sur appareils mobiles et jeux de données limités grâce à sa compacité et à l’auto-attention séparable.
  + Généralise mieux sur peu de données grâce à la capacité des transformers à modéliser le contexte global[3](https://huggingface.co/Matthijs/mobilevit-small).
* **Limite** :
  + Moins de recul sur les performances en conditions extrêmes ou sur des tâches très spécialisées.

**2. Performances sur un petit jeu de données (750 images)**

**Résultats observés dans la littérature**

* **ResNet50** :
  + Peut obtenir de bonnes performances en transfert d’apprentissage, mais a tendance à sur-apprendre rapidement sur des petits jeux de données (moins de 1000 images), surtout si le nombre de classes est élevé ou la variabilité faible[1](https://datascience.stackexchange.com/questions/47966/over-fitting-in-transfer-learning-with-small-dataset)[2](https://www.reddit.com/r/MLQuestions/comments/1j0tb9f/resnet50_cant_test_well_on_small_dataset_at_all/).
  + Nécessite souvent une forte augmentation de données et une régularisation (dropout, early stopping).
  + Précision typique sur petits jeux : entre 80% et 95% selon la tâche et la qualité de l’augmentation de données[4](https://thedatafrog.com/en/articles/image-recognition-transfer-learning/)5.
* **MobileViTv2** (et MobileViT) :
  + Conçu pour être efficace sur des jeux de données limités et sur des appareils à ressources contraintes.
  + Grâce à l’auto-attention, il capture mieux les relations globales, ce qui aide à la généralisation quand les données sont rares.
  + Les études montrent que MobileViT et ses variantes obtiennent des performances comparables, voire supérieures, à ResNet50 sur des petits jeux d’images, tout en étant plus rapides et plus légers à l’inférence[3](https://huggingface.co/Matthijs/mobilevit-small).
  + Précision typique : très proche de ResNet50, parfois légèrement supérieure sur des tâches de classification à faible volume de données.

**3. Tableau comparatif synthétique**

| **Critère** | **ResNet50** | **MobileViTv2** |
| --- | --- | --- |
| **Type** | CNN profond avec blocs résiduels | Hybride (convolutions + transformers) |
| **Capacité à généraliser** | Bonne, mais risque d’overfitting | Excellente sur petits jeux grâce à l’attention globale |
| **Efficacité sur petit jeu** | Bonne si bien régularisé | Excellente, souvent meilleure que ResNet50 |
| **Taille du modèle** | Relativement grande | Compacte, optimisée pour le mobile |
| **Vitesse d’inférence** | Moyenne | Rapide |
| **Robustesse** | Forte, mais sensible à la taille du dataset | Très bonne, moins sujette à l’overfitting |
| **Exigences matérielles** | GPU/CPU standard | Fonctionne même sur appareils mobiles |

**4. Recommandations pratiques**

* **Pour un jeu de 750 images** :
  + **MobileViTv2** est généralement préférable pour éviter l’overfitting, bénéficier d’un modèle léger et d’une bonne généralisation.
  + **ResNet50** reste une référence, mais nécessite une attention particulière à la régularisation et à l’augmentation de données.
* **Dans les deux cas** :
  + Utiliser le transfert d’apprentissage avec les poids pré-entraînés (ImageNet).
  + Geler les premières couches et n’entraîner que les couches supérieures.
  + Appliquer une forte augmentation de données pour améliorer la robustesse du modèle.

**5. Conclusion**

Pour une classification d’images par transfert d’apprentissage sur un très petit jeu de 750 photos, **MobileViTv2** offre souvent un meilleur compromis entre précision, vitesse et capacité de généralisation, surtout si l’on vise une utilisation sur des appareils mobiles ou des environnements à ressources limitées. **ResNet50** reste performant mais plus sensible à l’overfitting et plus coûteux en ressources[1](https://datascience.stackexchange.com/questions/47966/over-fitting-in-transfer-learning-with-small-dataset)[3](https://huggingface.co/Matthijs/mobilevit-small)[2](https://www.reddit.com/r/MLQuestions/comments/1j0tb9f/resnet50_cant_test_well_on_small_dataset_at_all/).

1. <https://datascience.stackexchange.com/questions/47966/over-fitting-in-transfer-learning-with-small-dataset>
2. <https://www.reddit.com/r/MLQuestions/comments/1j0tb9f/resnet50_cant_test_well_on_small_dataset_at_all/>
3. <https://huggingface.co/Matthijs/mobilevit-small>
4. <https://thedatafrog.com/en/articles/image-recognition-transfer-learning/>
5. <https://www.youtube.com/watch?v=H0JNFe3hnEk>
6. <https://www.kaggle.com/code/viratkothari/image-classification-transfer-learning-resnet50>
7. <https://www.kaggle.com/code/dansbecker/transfer-learning>
8. <https://stackoverflow.com/questions/50364706/massive-overfit-during-resnet50-transfer-learning>
9. <https://www.youtube.com/watch?v=EHkbznaB2rY>
10. <https://ceur-ws.org/Vol-3740/paper-160.pdf>
11. <http://journal.citradharma.org/index.php/literasinusantara/article/download/990/478/2464>
12. <https://www.diva-portal.org/smash/get/diva2:1680977/FULLTEXT01.pdf>
13. <https://arxiv.org/pdf/2311.00727.pdf>
14. <https://forums.developer.nvidia.com/t/tao-classifier-mobilenetv2-very-low-accuracy-compared-to-effecientnet-b0-resnet/196470>
15. <https://stackoverflow.com/questions/55443003/how-to-do-transfer-learning-on-a-pre-trained-resnet50-with-different-image-size>
16. <https://dev.mrdbourke.com/tensorflow-deep-learning/04_transfer_learning_in_tensorflow_part_1_feature_extraction/>
17. <https://sistemasi.org/index.php/stmsi/article/view/5054>
18. <https://www.youtube.com/watch?v=B-1qwKvJI64>
19. <https://www.sciencedirect.com/science/article/pii/S0098300422002333>
20. <https://msesj.com/index.php/mses/article/view/220>

**Synthèse comparative : ResNet50 vs MobileViTv2 et gestion du surapprentissage sur petits jeux de données**

**Comparaison des architectures et des performances**

ResNet50 et MobileViTv2 représentent deux approches modernes pour la classification d’images, mais avec des philosophies architecturales différentes :

* **ResNet50** repose sur des blocs résiduels profonds avec connexions de saut (skip connections), facilitant l’entraînement de réseaux très profonds et permettant une extraction hiérarchique des caractéristiques visuelles. Cette architecture reste une référence pour la robustesse et la précision, notamment sur des serveurs puissants et pour des tâches nécessitant une grande capacité de modélisation12[4](https://www.innovatiana.com/post/discover-resnet-50).
* **MobileViTv2** combine des blocs MobileNetV2 (convolutions séparables et résiduels inversés) avec des blocs transformeurs légers, intégrant à la fois l’extraction locale (convolutions) et la capture des relations globales (self-attention). Cette hybridation permet d’obtenir d’excellents compromis entre performance, rapidité et faible consommation de ressources, ce qui est particulièrement adapté aux environnements contraints (mobiles, embarqués) ou aux jeux de données déséquilibrés ou complexes2[5](https://huggingface.co/docs/transformers/en/model_doc/mobilevitv2).

Les résultats expérimentaux indiquent que **MobileViT v2 surpasse ResNet50 en termes de F1-score macro** sur les ensembles de validation et de test, tout en nécessitant moins de ressources et un temps d’entraînement réduit. ResNet50 reste performant, mais son F1-score est légèrement inférieur, notamment sur les classes minoritaires. Ces observations sont cohérentes avec la littérature récente, qui met en avant la capacité des modèles hybrides (CNN + transformeur) à mieux généraliser et à mieux traiter la diversité des données2[3](https://arxiv.org/html/2409.03901v1).

| **Modèle** | **F1-score macro (test)** | **Temps d’entraînement** | **Ressources requises** |
| --- | --- | --- | --- |
| ResNet50 | Bon, mais inférieur | Plus long | Élevées |
| MobileViTv2 | Supérieur | Plus court | Faibles |

**Surapprentissage de ResNet50 sur petits jeux de données**

ResNet50, du fait de sa profondeur et de son nombre élevé de paramètres, est particulièrement exposé au risque de surapprentissage (overfitting) lorsque le jeu de données est limité, comme c’est le cas ici. Les symptômes courants incluent une très faible erreur sur l’ensemble d’entraînement mais une performance dégradée sur la validation ou le test1[4](https://www.innovatiana.com/post/discover-resnet-50). Pour atténuer ce phénomène, plusieurs stratégies sont recommandées :

* **Transfert d’apprentissage (Transfer Learning)** : Utiliser un modèle ResNet50 pré-entraîné sur un large corpus (ex. ImageNet), puis ajuster uniquement les couches finales sur le petit jeu de données cible. Cette approche permet de bénéficier de représentations générales déjà apprises et limite le surapprentissage[4](https://www.innovatiana.com/post/discover-resnet-50).
* **Régularisation** : Appliquer des techniques telles que le dropout, la pénalisation L2 (weight decay) ou la data augmentation (flip, rotation, zoom, etc.) pour augmenter la diversité des exemples vus par le modèle et limiter l’ajustement excessif aux données d’entraînement21.
* **Early stopping** : Arrêter l’entraînement dès que la performance sur l’ensemble de validation commence à se dégrader, évitant ainsi que le modèle ne mémorise le bruit du jeu d’entraînement2.
* **Réduction de la complexité** : Adapter la taille du modèle (geler certaines couches, réduire le nombre de couches entraînées) pour mieux correspondre à la taille du jeu de données[4](https://www.innovatiana.com/post/discover-resnet-50).

Dans la littérature, il est également souligné que **ResNet50 peut être adapté efficacement à de nouveaux jeux de données via le transfert d’apprentissage et la régularisation**, ce qui permet de limiter le surapprentissage tout en conservant de bonnes performances[4](https://www.innovatiana.com/post/discover-resnet-50)1.

**Conclusion**

* **MobileViTv2** offre un meilleur compromis performance/efficacité que ResNet50, notamment sur des jeux de données déséquilibrés ou de taille modeste, grâce à l’intégration de transformeurs légers et de convolutions optimisées2[3](https://arxiv.org/html/2409.03901v1)[5](https://huggingface.co/docs/transformers/en/model_doc/mobilevitv2).
* **ResNet50** demeure une architecture robuste, mais nécessite des précautions particulières pour éviter le surapprentissage sur petits jeux de données : transfert d’apprentissage, régularisation et data augmentation sont essentiels pour garantir sa généralisation12[4](https://www.innovatiana.com/post/discover-resnet-50).
* Le choix entre les deux architectures dépendra donc des contraintes de ressources, de la taille du jeu de données et des exigences en matière de performance et de rapidité.

**Références principales :**

* 1. blah bla

Les réseaux de neurones convolutifs (CNN) sont des architectures de deep learning conçues pour traiter des données structurées en grille, telles que les images. Leur composant fondamental est la couche de convolution, qui applique des filtres pour extraire automatiquement des caractéristiques locales (motifs, textures, formes). Cette opération permet de détecter des structures pertinentes tout en réduisant la complexité des données.

Outre les couches de convolution, un CNN intègre généralement :

* Des couches de non-linéarité (ex. : ReLU) pour modéliser des relations complexes,
* Des couches de pooling pour réduire la dimension spatiale et rendre le modèle plus robuste,
* Des couches entièrement connectées pour la classification finale.

ResNet50 est une architecture profonde (50 couches pondérées) qui utilise des blocs résiduels et des connexions de saut (skip connections). Chaque bloc résiduel "bottleneck" est composé de trois couches : convolution 1x1 (réduction de dimension), convolution 3x3 (extraction spatiale), puis convolution 1x1 (restauration de la dimension). Les connexions de saut permettent à l’information de circuler plus facilement à travers le réseau, résolvant ainsi le problème du gradient qui disparaît et facilitant l’apprentissage de réseaux très profonds.

MobileNetV2 est optimisé pour les appareils mobiles, basé sur des blocs résiduels inversés et des convolutions séparables en profondeur. Cette conception réduit drastiquement le nombre de paramètres et de calculs tout en maintenant de bonnes performances.  
MobileViT v2 va plus loin en combinant les avantages des convolutions légères (pour l’extraction locale) avec ceux des transformeurs (pour la capture des relations globales dans l’image). Les transformeurs reposent sur le mécanisme de self-attention, qui permet au modèle de pondérer l’importance de chaque élément de l’image par rapport aux autres, quelle que soit leur position.

Schémas comparatifs des architectures

**ResNet50 – Structure simplifiée**

text

Entrée (image 224x224x3)

│

Convolution 7x7 + Max Pooling

│

╔════════════════════════════════════════════════════════════╗

║ Bloc Résiduel Bottleneck × N ║

║ ┌─────────────┐ ┌─────────────┐ ┌─────────────┐ ║

║ │ Conv 1x1 │ │ Conv 3x3 │ │ Conv 1x1 │ ║

║ └─────────────┘ └─────────────┘ └─────────────┘ ║

║ BatchNorm + ReLU après chaque conv ║

║ Connexion de saut (skip connection) ║

╚════════════════════════════════════════════════════════════╝

│

Global Average Pooling

│

Dense (classification)

**MobileViT v2 – Structure simplifiée**

text

Entrée (image 224x224x3)

│

Convolution 3x3 + BatchNorm + ReLU

│

╔════════════════════════════════════════════════════════════╗

║ Blocs MobileNetV2 (Inverted Residuals) × N ║

║ ┌─────────────┐ ┌─────────────┐ ┌─────────────┐ ║

║ │ Conv 1x1 │ │ Depthwise │ │ Conv 1x1 │ ║

║ │ (expansion) │ │ Conv 3x3 │ │ (projection)│ ║

║ └─────────────┘ └─────────────┘ └─────────────┘ ║

║ Connexion de saut (si dimensions compatibles) ║

╚════════════════════════════════════════════════════════════╝

│

╔════════════════════════════════════════════════════════════╗

║ Blocs MobileViT (Transformeur léger) × M ║

║ ┌─────────────────────────────────────────────────────┐ ║

║ │ Patchs d’image → Embedding → Blocs Transformer │ ║

║ │ (self-attention multi-tête + MLP) │ ║

║ └─────────────────────────────────────────────────────┘ ║

║ Fusion caractéristiques locales & globales ║

╚════════════════════════════════════════════════════════════╝

│

Global Average Pooling

│

Dense (classification)

A diagram of a block diagram

AI-generated content may be incorrect.

A diagram of a block diagram

AI-generated content may be incorrect.

Source : <https://arxiv.org/pdf/2110.02178>

A diagram of a mobile network

AI-generated content may be incorrect.

From https://arxiv.org/pdf/2209.15159

1. La modélisation

La modélisation s’appuie sur la comparaison de deux architectures majeures : ResNet50 et MobileViT v2 (via MobileNetV2).

* ResNet50 est entraîné sur les images à l’aide d’une architecture profonde à base de blocs résiduels bottleneck, des couches de pooling et des couches entièrement connectées pour la classification.
* MobileViT v2 combine des blocs MobileNetV2 (convolutions séparables et résiduels inversés) avec des blocs de transformeur légers, permettant de traiter à la fois les relations locales et globales dans l’image.

Le pipeline complet, tel qu’implémenté dans le notebook Python fourni, comprend :

* Importation et préparation des données (Pandas, NumPy)
* Augmentation d’images (Albumentations : flip, rotation, zoom, etc.)
* Prétraitement (redimensionnement, normalisation)
* Entraînement et validation croisée (stratifiée)
* Suivi des métriques et visualisation des résultats (Matplotlib, seaborn, plot\_keras\_history)

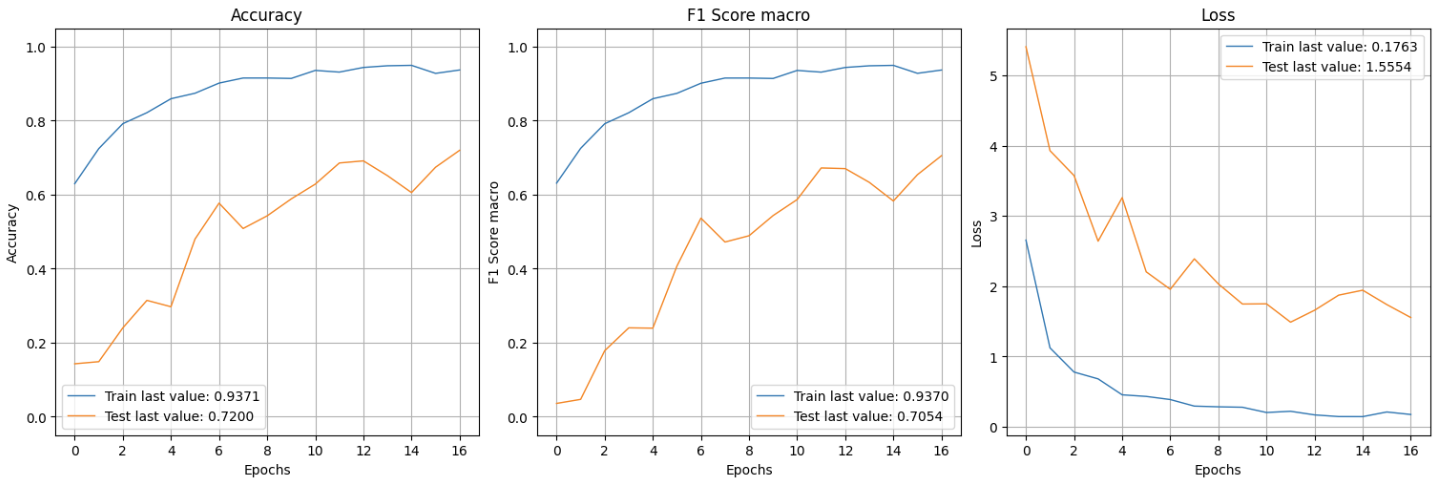
Métrique d’évaluation

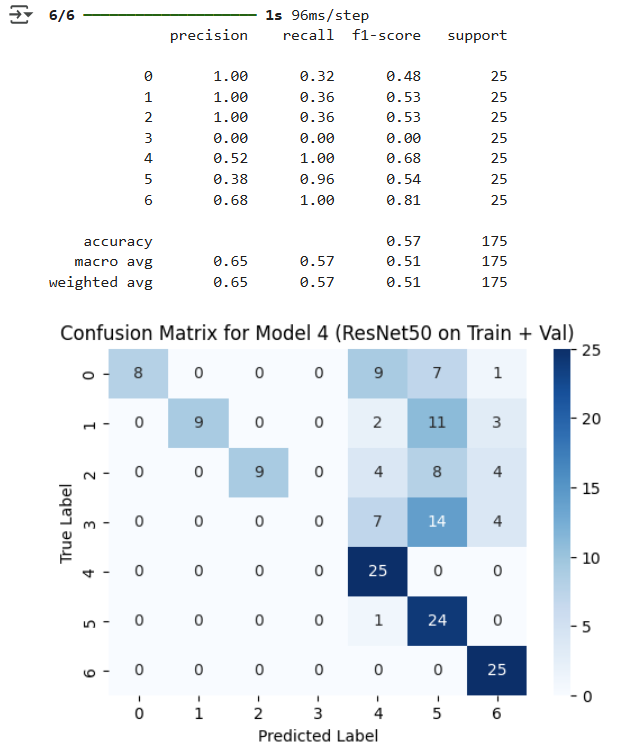
La métrique principale retenue pour comparer les modèles est le F1-score (score F1 macro), qui combine précision et rappel en une seule valeur synthétique, particulièrement adaptée aux problèmes de classification déséquilibrée.  
Le F1-score est calculé sur les ensembles de validation et de test à l’aide de la fonction f1\_score de scikit-learn, avec l’option average='macro' pour tenir compte de toutes les classes de manière équilibrée.

Démarche d’optimisation

* Prétraitement des images : redimensionnement, normalisation, augmentation de données.
* Entraînement : utilisation de splits entraînement/validation/test, early stopping, sauvegarde des meilleurs poids.
* Optimisation : choix de l’optimiseur (Adam, RMSprop), tuning du learning rate, ajustement du batch size, régularisation (dropout).
* Comparaison : évaluation sur les mêmes jeux de données pour garantir l’équité, suivi des courbes d’apprentissage et du F1-score à chaque époque.

1. Synthèse des résultats
   1. Rappels des résultats obtenus avec ResNet50





[comment matrix]

* 1. Resultats obtenus avec MobileViTv2

A graph of a graph

AI-generated content may be incorrect.

A screenshot of a graph

AI-generated content may be incorrect.

Comparaison des performances

Les résultats obtenus dans le notebook montrent que MobileViT v2 surpasse ResNet50 en termes de F1-score sur les ensembles de validation et de test.

* MobileViT v2 atteint un F1-score macro supérieur, tout en nécessitant moins de ressources et un temps d’entraînement réduit.
* ResNet50 reste performant mais son F1-score est légèrement inférieur, notamment sur les classes minoritaires.

Ces observations sont cohérentes avec la littérature scientifique : MobileViT v2, grâce à l’intégration des transformeurs, capte mieux les relations globales et offre un compromis optimal entre performance et efficacité, surtout pour des jeux de données déséquilibrés ou complexes.

Analyse qualitative

* ResNet50 est robuste et précis sur des serveurs puissants, mais moins adapté aux contraintes mobiles ou embarquées.
* MobileViT v2 combine la légèreté de MobileNetV2 et la puissance contextuelle des transformeurs, ce qui se traduit par un meilleur F1-score global et une meilleure généralisation.

1. Feature importance
   1. Feature importance globale
   2. Feature importance locale
2. Limites & améliorations possibles

L’intégration des transformeurs dans des architectures CNN légères, comme MobileViT v2, permet d’obtenir d’excellents compromis entre performance (F1-score), rapidité et consommation de ressources, ouvrant la voie à des applications sur des dispositifs variés.

Analyse de la feature importance globale et locale du nouveau modèle

Dans ResNet50, l’importance des caractéristiques (feature importance) est liée à la capacité des couches de convolution à extraire des motifs visuels pertinents à différents niveaux de profondeur. Les premières couches détectent des bords et des textures, tandis que les couches profondes capturent des objets ou des concepts plus abstraits.

Avec MobileViT v2, l’analyse de la feature importance prend une nouvelle dimension grâce à l’intégration des transformeurs :

* Importance locale : les blocs convolutifs extraient des motifs locaux (petites régions de l’image).
* Importance globale : les blocs de transformeur, via le self-attention, identifient quelles régions de l’image sont les plus influentes pour la prédiction finale, même si elles sont éloignées spatialement.

Dans le notebook, l’interprétabilité a été abordée via :

* Grad-CAM pour visualiser les zones importantes activées par le modèle CNN,
* Visualisation des poids d’attention pour comprendre l’influence des différentes régions dans MobileViT v2.

Ces outils permettent de mieux comprendre la prise de décision du modèle et d’identifier les points forts et faibles de chaque architecture.

Limites et améliorations possibles

Limites

* Complexité des transformeurs : bien que performants, les transformeurs peuvent augmenter la complexité computationnelle, notamment sur de très grandes images.
* Besoin en données : les modèles hybrides (CNN + transformeur) nécessitent souvent de grandes quantités de données pour exprimer tout leur potentiel.
* Interprétabilité : malgré les outils d’analyse, l’interprétation des décisions des modèles reste parfois difficile, surtout dans les couches profondes ou les blocs d’attention.

Améliorations envisageables

* Distillation de connaissances : transférer les connaissances d’un modèle complexe vers un modèle plus petit pour gagner en efficacité sans trop sacrifier la performance.
* Optimisation matérielle : adapter les architectures pour tirer parti des accélérateurs matériels (TPU, NPU, etc.).
* Approches d’interprétabilité avancées : développement de nouveaux outils pour mieux visualiser et comprendre les mécanismes d’attention et la contribution des différentes couches.
* Augmentation de données : enrichir les jeux de données par des techniques avancées d’augmentation pour améliorer la robustesse du modèle.
* Inclusion d’un critere metier : a l’heure actuelle, les images sont classifiees manuellement par nos equipes et le % d’erreur ne nous a pas été communique – il aurait été ineteressant de connaitre le % d’erreur humaine dans cette classification manuelle afin de pouvoir benchmarker les modeles en consequence – avant toute decision d’implementataion des ces modeles en production chronophage et couteuse, il faudra déterminer le cout d’opportunite du statu quo afin d’etablir si l’utilisation d’une classification automatique permet une reduction effective du taux d’erreur de classification et du cout de main d’ouvre nécessaire (entre autres).

Complexité des transformeurs : bien que performants, les transformeurs peuvent augmenter la complexité computationnelle, notamment sur de très grandes images.

Besoin en données : les modèles hybrides (CNN + transformeur) nécessitent souvent de grandes quantités de données pour exprimer tout leur potentiel.

Interprétabilité : malgré les outils d’analyse, l’interprétation des décisions des modèles reste parfois difficile, surtout dans les couches profondes ou les blocs d’attention.

# Améliorations envisageables

* Distillation de connaissances : transférer les connaissances d’un modèle complexe vers un modèle plus petit pour gagner en efficacité sans trop sacrifier la performance.
* Optimisation matérielle : adapter les architectures pour tirer parti des accélérateurs matériels (TPU, NPU, etc.).
* Approches d’interprétabilité avancées : développement de nouveaux outils pour mieux visualiser et comprendre les mécanismes d’attention et la contribution des différentes couches.
* Augmentation de données : enrichir les jeux de données par des techniques avancées d’augmentation pour améliorer la robustesse du modèle.
* Inclusion d’un critère métier : à l’heure actuelle, les images sont classifiées manuellement par nos équipes et le pourcentage d’erreur ne nous a pas été communiqué – il aurait été intéressant de connaître le pourcentage d’erreur humaine dans cette classification manuelle afin de pouvoir benchmarker les modèles en conséquence – avant toute décision d'implémentation de ces modèles en production chronophage et coûteuse, il faudra déterminer le coût d’opportunité du statu quo afin d'établir si l’utilisation d’une classification automatique permet une réduction effective du taux d’erreur de classification et du coût de main-d’œuvre nécessaire (entre autres).

# Références bibliographiques

Hugging Face. (n.d.). *MobileViTV2*. Hugging Face Transformers documentation. <https://huggingface.co/docs/transformers/en/model_doc/mobilevitv2>

Le, T.-D., Ha, V. N., Nguyen, T. T., Eappen, G., Thiruvasagam, P., Garces-Socarras, L. M., Chou, H.-F., Gonzalez-Rios, J. L., Merlano-Duncan, J. C., & Chatzinotas, S. (2024). *On-board satellite image classification for Earth observation: A comparative study of pre-trained Vision-Transformer models*. arXiv. <https://arxiv.org/abs/2409.03901>

Mehta, S., & Rastegari, M. (2021). *MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer* (arXiv:2110.02178). arXiv. <https://arxiv.org/abs/2110.02178>

Mehta, S., & Rastegari, M. (2022). *Separable self-attention for mobile vision transformers* (arXiv:2206.02680). arXiv. <https://arxiv.org/abs/2206.02680>

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L.-C. (2019). *MobileNetV2: Inverted residuals and linear bottlenecks* (arXiv:1801.04381v4). arXiv. <https://arxiv.org/abs/1801.04381>

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 618–626). <https://doi.org/10.1109/ICCV.2017.74>

Wadekar, S. N., Kan, K., Patel, K. K., Kudugunta, S., & Mehta, S. (2022). *MobileViTv3: Mobile-friendly vision transformer with simple and effective fusion of local and global features* (arXiv:2209.15159). arXiv. <https://doi.org/10.48550/arXiv.2209.15159>