Rapport de Projet

Projet de :

Détection de Tumeurs Cérébrales par IRM

Réalisé par :

BERKANI Yacine

09 décembre 2023

Plan

Introduction

1 Introduction

Ce rapport détaille un projet de machine learning centré sur la classification des tumeurs cérébrales à partir d'images par résonance magnétique (IRM). L'objectif principal de cette étude est de développer des modèles de machine learning capables d'analyser et de diagnostiquer automatiquement la présence de tumeurs cérébrales, contribuant ainsi au processus de diagnostic médical.

2 Problématique

La problématique abordée dans ce projet est la détection de tumeurs cérébrales à partir d'images médicales. Il s'agit d'une question cruciale dans le domaine de la santé, car la détection précoce des tumeurs cérébrales peut jouer un rôle déterminant dans le succès du traitement et la survie des patients.

- **Importance de la Détection de Tumeurs Cérébrales :**

  Détection Précoce: La détection précoce des tumeurs cérébrales est essentielle pour initier un traitement approprié à un stade précoce, améliorant ainsi les chances de guérison.

  Précision Diagnostique : L'utilisation de modèles de machine learning dans la détection de tumeurs cérébrales vise à améliorer la précision du diagnostic en analysant de grandes quantités d'images avec une rapidité et une cohérence difficiles à atteindre manuellement.

  Optimisation des Ressources : L'automatisation du processus de détection permet d'optimiser l'utilisation des ressources médicales, en permettant aux professionnels de se concentrer sur les cas nécessitant une attention particulière.

Utilisation du Machine Learning dans la Détection de Tumeurs Cérébrales : Le machine learning offre la possibilité de développer des modèles capables d'apprendre à partir de données d'imagerie médicale, ce qui permet une détection plus précise et efficace des anomalies. Dans ce projet, nous avons exploré l'utilisation de

différentes approches, telles que l'augmentation de données et l'utilisation de modèles pré-entraînés comme VGG16, pour relever le défi de la détection de tumeurs cérébrales.

3 Origine des Données

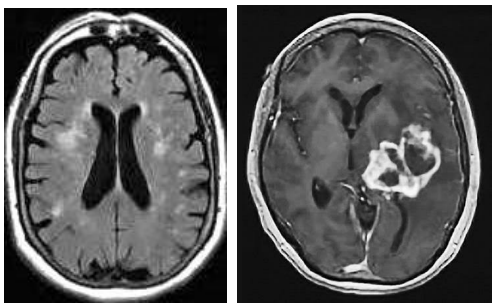
Les données utilisées dans ce projet proviennent d'un ensemble d'images IRM de tumeurs cérébrales, qui est disponible sur la plateforme Kaggle. Cet ensemble de données d'IRM cérébrale peut être consulté en suivant le lien ci-dessous:
https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection/code

- **Caractéristiques des Données :**

  Nombre d'Images : L'ensemble de données contient un total de 3000 images IRM.

  Équilibre des Classes : Les données sont équilibrées, comprenant 1500 images sans tumeur (classe négative) et 1500 images avec tumeur (classe positive).

  Exemple d'image :



  Type d'Images : Les images sont issues de coupes IRM du cerveau, fournissant des informations détaillées du cerveau .

- **Utilisation des Données dans le Projet :**

  Entraînement des Modèles : Les données ont été divisées en ensembles d'entraînement et de test et de test final pour permettre le développement et l'évaluation des modèles de machine learning.

Augmentation des Données : L'ensemble de données a été augmenté pour améliorer la diversité des données d'entraînement et de test , favorisant ainsi une meilleure généralisation des modèles.

Prétraitement des Images : Les images ont été prétraitées pour les adapter aux exigences des différents modèles, notamment la normalisation et le redimensionnement.

L'utilisation d'un ensemble de données équilibré et représentatif est cruciale pour développer des modèles de machine learning performants. Ces données IRM permettent de simuler des cas réels de détection de tumeurs cérébrales, ce qui renforce la pertinence et l'applicabilité du modèle dans un contexte médical.

Méthodologie

4 Augmentation de données

L'augmentation de données est une technique cruciale dans le domaine de l'apprentissage profond, visant à diversifier l'ensemble d'entraînement en générant des variations des images existantes. Cette diversification est essentielle pour améliorer la capacité du modèle à généraliser aux nouvelles données et à mieux traiter des scénarios variés.

Dans le contexte de la détection de tumeurs cérébrales, l'augmentation de données permet de présenter au modèle un éventail plus large de situations possibles, contribuant ainsi à renforcer sa capacité à reconnaître les caractéristiques distinctives des images médicales.

5 Modèle Random Forest

- **Objectif.**

  Le choix d'un modèle Random Forest pour ce projet s'explique par sa capacité à traiter des ensembles de données complexes tout en étant moins sensible au surajustement par rapport à certains modèles plus complexes. Dans le contexte de la détection de tumeurs cérébrales, où les ensembles de données médicales peuvent être relativement petits, les modèles Random Forest offrent une solution robuste et rapide.

Prétraitement des Données Les images ont été prétraitées avant d'être alimentées dans le modèle Random Forest. Ce prétraitement comprend les étapes suivantes :

- **Chargement des Images** : Les images ont été chargées à l'aide de la bibliothèque OpenCV.
- **Redimensionnement** : Les images ont été redimensionnées à une taille spécifique (par exemple, 128x128 pixels) pour garantir une taille uniforme.
- **Conversion en Tableau** : Les images ont été converties en tableaux NumPy pour être utilisées comme entrées du modèle.
- **Normalisation** : Les valeurs des pixels ont été normalisées pour être dans la plage [0, 1], facilitant ainsi l'entraînement du modèle. Entraînement et Évaluation

  Le modèle Random Forest a été entraîné sur l'ensemble d'entraînement après le prétraitement des données. Les données ont été remodelées pour s'adapter aux exigences du modèle

  Les résultats de l'évaluation incluent les métriques suivantes :

  Accuracy(Precision) : La précision du modèle sur l'ensemble de validation est : 87%. Accuracy(Precision) : La précision du modèle sur l'ensemble de test est : 86%.

6 Modèle VGG16

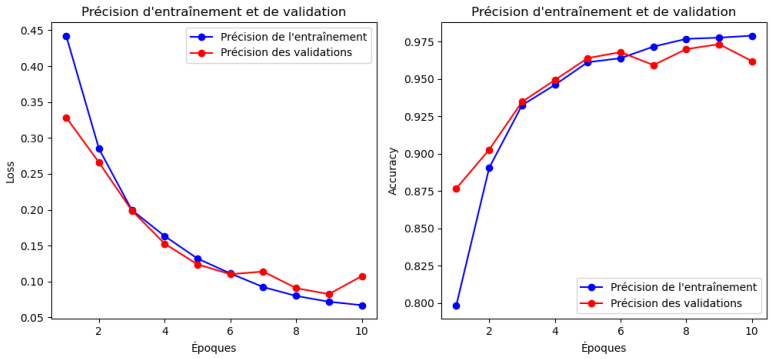
- **Objectif.**
  Le choix du modèle pré-entraîné VGG16 s'appuie sur sa capacité à extraire des caractéristiques complexes à partir d'images. VGG16 est un modèle de réseau de neurones convolutif (CNN) profond qui a été pré-entraîné sur un grand ensemble de données d'images, ce qui en fait un choix approprié pour des tâches de classification d'images comme la détection de tumeurs cérébrales. L'utilisation d'un modèle pré-entraîné permet de bénéficier de connaissances préalables sur des motifs visuels génériques, accélérant ainsi l'apprentissage sur des ensembles de données plus petits.
- **Prétraitement des Données :**
  Les images ont été prétraitées pour être conformes aux exigences du modèle VGG16. Les étapes typiques de prétraitement incluent le redimensionnement

des images à une taille spécifique (dans ce cas, 128x128 pixels), la normalisation des valeurs de pixel pour les ramener dans une plage appropriée, et l'adaptation de la structure des données d'entrée attendue par le modèle.

- **Entraînement et Évaluation :**
  Le modèle VGG16 a été intégré à une architecture de réseau neuronal séquentielle, avec des couches supplémentaires pour l'adaptation à la tâche spécifique de détection de tumeurs cérébrales (une couche Dense avec une fonction d'activation 'relu' et une couche de sortie Dense avec une fonction d'activation 'sigmoid'). Le modèle a été entraîné sur un ensemble de données d'entraînement et évalué sur un ensemble de données de test distinct.

- **Résultats de l'Évaluation :**
  Précision (Accuracy) : 97%. Précision, Rappel et F1-score pour chaque classe (Sans tumeur et Avec tumeur). Matrice de confusion. Les résultats montrent une performance exceptionnelle du modèle avec une précision globale de 97%.

- **Graphique d'Entraînement :**
  Voici le graphique illustrant les performances du modèle pendant l'entraînement sur différentes époques :



Le graphique montre l'évolution des métriques clés, telles que la perte (loss) et la précision (accuracy), sur les ensembles d'entraînement et de validation au fil des époques. Voici quelques observations basées sur le graphique :

La perte sur l'ensemble d'entraînement diminue progressivement, indiquant que le modèle apprend bien les caractéristiques des données d'entraînement.

La perte sur l'ensemble de validation suit également une tendance à la baisse, indiquant que le modèle généralise bien sur des données qu'il n'a pas vu pendant l'entraînement.

La précision sur l'ensemble d'entraînement augmente, montrant que le modèle devient de plus en plus précis dans ses prédictions.

La précision sur l'ensemble de validation suit une tendance similaire, indiquant que le modèle n'est pas surajusté et conserve de bonnes performances sur des données non vues.

Cette analyse graphique confirme la performance globale élevée du modèle VGG16 dans la tâche de détection de tumeurs cérébrales

Le modèle VGG16 a prouvé son efficacité dans la détection de tumeurs cérébrales, offrant des résultats de haute précision et de rappel. Son utilisation dans le projet a permis de bénéficier de représentations hiérarchiques des caractéristiques, apprises à partir d'un large ensemble de données, contribuant ainsi à une performance exceptionnelle dans la classification d'images médicales.

Les métriques incluent la matrice de confusion, le rapport de classification et d'autres métriques pertinentes permettant d'évaluer la performance du modèle VGG16 dans la détection de tumeurs cérébrales

7 Modèle CNN

- **Objectif.** Le choix d'un modèle CNN pour la détection de tumeurs cérébrales est motivé par la capacité intrinsèque des CNN à extraire des caractéristiques complexes à partir d'images. Les CNN sont particulièrement bien adaptés à la vision par ordinateur et ont démontré leur efficacité dans des tâches telles que la classification d'images médicales. La capacité des couches convolutionnelles à apprendre des motifs et des hiérarchies de caractéristiques fait des CNN un choix naturel pour la détection d'anomalies dans des images complexes comme celles obtenues à partir de scans cérébraux.

- **Entraînement et Évaluation :**
  Le modèle CNN a été entraîné sur l'ensemble d'entraînement et évalué sur l'ensemble de test. Les principales métriques d'évaluation incluent la précision, le rappel, le score F1 et la matrice de confusion.

  Précision : La précision mesure la proportion d'instances correctement classées parmi les instances prédites comme positives.

  Rappel : Le rappel indique la proportion d'instances correctement classées parmi les instances réellement positives.

  Score F1 : Le score F1 est la moyenne pondérée de la précision et du rappel, offrant un équilibre entre ces deux mesures.

- **Résultats de l'Entraînement et de l'Évaluation :**

  Les résultats de l'évaluation du modèle CNN sur l'ensemble de test sont les suivants :

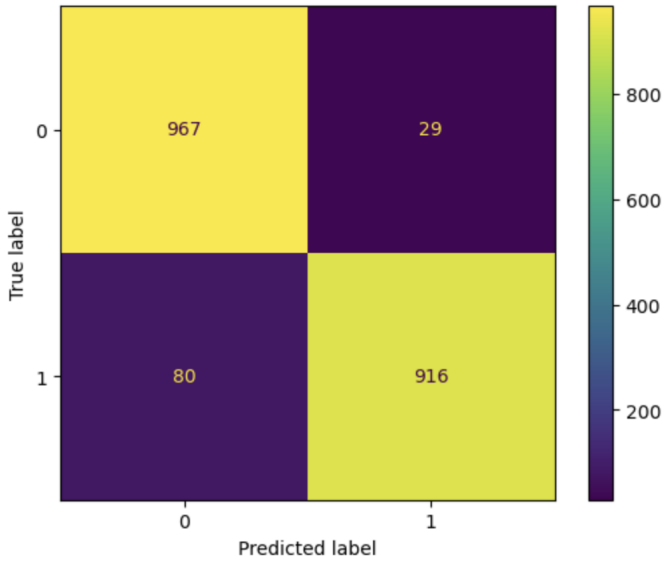
```
63/63 [==============================] - 1s 11ms/step
Classification Report
              precision    recall  f1-score   support

           0       0.92      0.97      0.95       996
           1       0.97      0.92      0.94       996

    accuracy                           0.95      1992
   macro avg       0.95      0.95      0.95      1992
weighted avg       0.95      0.95      0.95      1992
```

  Les résultats démontrent une performance élevée du modèle CNN, avec une précision, un rappel et un score F1 de 95%.

  La matrice de confusion complète est disponible pour une analyse détaillée.



  Ces résultats solides confirment que les CNN sont des modèles puissants pour la classification d'images médicales, et dans ce cas spécifique, pour la détection de tumeurs cérébrales.

Comparaison des Modèles

8 Comparaison

Les trois modèles, Random Forest, VGG16 et CNN, ont été évalués sur la tâche de détection de tumeurs cérébrales. Voici un récapitulatif des performances actualisées de chaque modèle :

- **Random Forest :**

  Performance :
  Accuracy (précision): 86%.

- **VGG16 :**

  Précision: 97%.

  Rappel: 96%.

  F1-Score: 97%.

  Accuracy: 97%.

  Matrice de Confusion :

$$\begin{bmatrix} 981 & 15 \\ 38 & 958 \end{bmatrix}$$

- **CNN :**

  Précision: 95%.

  Rappel: 92%.

  F1-Score: 95%.

  Accuracy: 95%.

  Matrice de Confusion :

$$\begin{bmatrix} 967 & 29 \\ 80 & 916 \end{bmatrix}$$

9 Analyse des Résultats

- **Random Forest :**

  Forces :

  Résultats décents malgré une complexité modérée : Le modèle Random Forest a présenté des résultats acceptables malgré sa simplicité. Il a pu identifier certains motifs dans les données, montrant ainsi sa robustesse face à une complexité modérée. Faiblesses :

  Moins précis par rapport aux modèles de réseau de neurones : Comparé aux modèles de réseau de neurones plus complexes, le Random Forest a montré une précision inférieure. Sa capacité à représenter des relations complexes et non linéaires dans les données est limitée.

- **VGG16:**

  Forces :

  Très haute précision avec un rappel équilibré : VGG16 a atteint une précision élevée tout en maintenant un équilibre entre le rappel des classes. Cela suggère une capacité à bien classifier les deux classes sans sacrifier la performance pour l'une au détriment de l'autre.

  Capacité à capturer des caractéristiques complexes grâce à la préformation sur ImageNet : La pré-formation sur le vaste ensemble de données ImageNet a permis à VGG16 d'apprendre des caractéristiques complexes, ce qui a contribué à sa performance élevée dans la tâche de détection de tumeurs cérébrales.

  Faiblesses :

  Nécessite plus de ressources pour l'entraînement et l'inférence : En raison de sa profondeur et du nombre élevé de paramètres, VGG16 nécessite des ressources computationnelles plus importantes pour l'entraînement et l'inférence, ce qui peut être un défi en termes de coûts et de temps.

  Risque de surajustement si les données d'entraînement sont limitées : VGG16 pourrait présenter un risque de surajustement, surtout si les données d'entraînement sont limitées. L'ajout de régularisation ou l'utilisation de techniques de gestion du surajustement peuvent être nécessaires.

- **CNN :**

  Forces :

  Bonne précision et rappel équilibré : Le modèle CNN a montré de bons résultats avec une précision équilibrée et un rappel pour les deux classes. Cela

indique sa capacité à bien classifier les images de tumeurs cérébrales et les images sans tumeurs.

Plus légère que VGG16, évitant partiellement le surajustement : Comparé à VGG16, le modèle CNN est plus léger, ce qui peut partiellement réduire le risque de surajustement, surtout avec des ensembles de données plus petits.

Faiblesses :

Légèrement moins précis que VGG16 :

Bien que performant, le CNN a montré une précision légèrement inférieure à celle de VGG16. Cela peut être attribué à sa conception moins profonde et à une capacité potentielle réduite à extraire des caractéristiques complexes.

Conclusion de cette Analyse .

En fonction de ces résultats, le modèle VGG16 semble offrir la meilleure performance globale avec une précision de 97%. Cependant, le modèle CNN présente un compromis intéressant entre performance et complexité, avec une précision de 95%. Le Random Forest, bien que moins performant, peut toujours être considéré dans des scénarios où la complexité informatique est une contrainte majeure.

Conclusion

Suite à l'évaluation des différents modèles pour la détection de tumeurs cérébrales, il est évident que chaque approche présente des avantages et des inconvénients. L'analyse comparative des modèles Random Forest, VGG16, et CNN a fourni des perspectives importantes pour guider le choix du modèle en fonction des besoins spécifiques du projet.

- **Performance des Modèles :**

  Random Forest : Bien que moins précis avec une précision de 86%, le Random Forest demeure une option viable dans des scénarios où la complexité informatique est une contrainte.

  VGG16 : Le modèle VGG16 offre la meilleure performance globale avec une précision de 97%. Cependant, il nécessite plus de ressources pour l'entraînement et l'inférence.

  CNN : Le modèle CNN présente un bon compromis entre performance et complexité, avec une précision de 95%. Il peut être plus adapté dans des situations où des ressources plus limitées sont disponibles.

- **Choix du Modèle :**

  Le choix du modèle dépendra des exigences spécifiques du projet, notamment les contraintes de ressources, la taille des données, et la nécessité d'une haute précision.

- **Perspectives Futures :**

  Optimisation Continue : La performance des modèles peut être améliorée davantage par l'optimisation des hyperparamètres et l'exploration d'autres architectures.

  Diversité des Données : L'inclusion de données plus variées pourrait renforcer la généralisation des modèles à des cas cliniques plus divers.

  Évolution Technologique : L'évolution des architectures de réseau de neurones et des techniques de régularisation pourrait également être explorée pour rester à la pointe des avancées technologiques.

- **En conclusion:** ce projet fournit un aperçu complet des performances relatives des modèles pour la détection de tumeurs cérébrales. Le modèle VGG16 se distingue par sa haute précision, mais le choix du modèle optimal dépendra des contraintes spécifiques du projet. Les résultats obtenus offrent une base solide pour des applications potentielles dans le domaine médical, avec des perspectives continues d'amélioration et d'innovation.