La compression des réseaux de neurones profonds à convolution pour l'analyse des images cardiaquescoronariennes

Convolutional deep neural network compression for Coronarian cardiac image analysis

Roufaida Trad, Nabiha Azizi*, Assia Boukhamla
Laboratoire LABGED, Département d'Informatique, Faculté de Technologie,
Université Badji Mokhtar, Annaba 23005, Algeria.

Info. Article

Historique de l'article

Received 01/08/2022
Revised 02 /10/2022
Accepted 02/10/2022

Mots-clés :

Réseaux de neurones profonds, VGG16, compression, maladie coronarienne, quantification, normalisation, transforméeenondelettediscrète

RESUME

Le travail présenté aborde l'optimisation des modèles profonds basée compression pour laclassification des images médicales plus particulièrement les images cardiaques. Nous distinguons deux types de compressions, celle des modèles et celle des images. Le modèle d'apprentissage profond simplifié que nous ciblons est un modèle dont la taille et le temps de réponses ont été réduites par rapport à l'original sans pour autant diminuer considérablement la précision initiale. L'objectif de ce travail est de tester l'aptitude de ces deux stratégies de compression afin de résoudre le problème initialement posé. L'investigation des modalities de compression est établie pour l'optimisation d'un système d'aide au diagnostic médical (maladie coronarienne), à travers un modèle covolutionnel par transfer learning basé sur le classifieur VGG16. Validé sur la base médicale référenicée "*CAD Cardiac MRI Dataset*", les results montrent que l'utilisation de la quantification assure une réduction significative de la taille des modèles et une optimisation remarquable en terme de précision et temps d'inférence. Ceci garanti le déploiement du complexe modèles profonds sur des périphiriques à ressources limitées tout en gardant la performance de décision du modèle non quantifié.

* **Auteur Correspondant :**
Email: azizi@labged.net

1. INTRODUCTION

Les systèmes d'aide à la décision médicale connaissent une évolution rapide et leur généralisation auprès des médecins se fait d'une façon croissante. En effet, les cliniciens sont de plus en plus ouverts à profiter de l'assistance des nouvelles percées technologiques et exploiter tout outil qui pourrait leur faciliter la tâche et notamment leur permettre d'améliorer la précision de leurs diagnostics et l'efficacité des traitements préconisés.

La maladie des artères coronaires est la première cause de mortalité dans le monde, selon l'organisation mondiale de la santé, elle fait référence au rétrécissement des artères qui fournissent du sang au cœur provoqué par l'accumulation de dépôts graisseux sur la paroi des artères. Comme pour toute maladie à fort taux d'incidence ou de mortalité, le dépistage et le diagnostic sont d'un grand intérêt de santé publique donc il est important de développer un système d'aide au diagnostic propre à cette maladie.

Ces dernières années les réseaux de neurones profonds ou Deep Neural Networks (DNNs) ont atteint des performances inégalées en simplifiant plusieurs tâches difficiles dans les domaines de la vision par ordinateur, de la reconnaissance vocale et du langage, etc [1,2]. Les données suffisamment volumineuses et la puissance de calcul avancée permettent l'utilisation de réseaux de neurones avec des architectures plus larges et plus profondes qui offrent des performances optimales dans de nombreuses applications. Cependant, cela s'accompagne par une taille des architectures d'apprentissage profond plus importante : les réseaux de neurones profonds modernes peuvent être composés de centaines de millions de paramètres, ce qui les rend difficiles à stocker et lents à former et pose ainsi problème pour le déploiement des modèles d'apprentissage profond sur des appareils à ressources limitées (par exemple, téléphones mobiles, robots et microcontrôleurs).

Pour faire face à la contrainte liée aux ressources limitées, l'exploitation des techniques de compression s'avère une approche indispensable. On distingue deux types de compressions. La compression des modèles qui permet à optimiser les DNNs en diminuant leurs tailles et en réduisant les coûts de calcul et les besoins de stockage tout en concervant la précision du modèle, et la compression des images qui consiste à minimiser la taille en octets d'un fichier graphique sans dégrader la qualité de l'image à un niveau inacceptable.

L'objectif principal de notre travail est à explorer ces deux méthodes de compression, appliquées aux systèmes d'aide au diagnostic propre à la maladie coronarienne.

2. TRAVAUX CONNEXES

Plusieurs études de recherche récentes explorent les méthodes ML et DL pour le diagnostic de CAD. En ce qui concerne l'application de DL, Berkaya, Sivrikoz et Gunal (2020) [3] ont proposé deux modèles de classification différents pour la classification des images SPECT, à savoir, le premierest basé sur l'apprentissage en profondeur (DL) tandis que la'autre est basé sur les connaissances. Les performances atteintes étaient au voisinage de 94 % pour le modèle basé sur DL et à 93 % pour le modèle basé sur les connaissances, respectivement. Le meilleur modèle DL a été déterminé comme étant VGG16 avec des fonctionnalités profondes de la machine à vecteurs de support (SVM) peu profondes. Nikolaos et Elpiniki (2021) [4] ont proposé un modèle CNN basé RGB pour identifier la catégorie de diagnostic CA. En présentant ses capacités de généralisation, les résultats ont révélé une précision globale de la classification de 93,47 % en utilisant des scans SPECT MPI.

En raison des exigences de calcul massives des modèles CNN récents, le besoin de méthodes d'optimisation des modèles était une obligation à prendre en considération. Les auteurs [5] ont évalué 200 images ultrasonic pour le diagnostic du cancer du sein en testant 3 réseaux de neurones profonds : VGG16, Googlenet et Resnet34 avec 3 méthodes d'optimisation de quantification. Cet article a démontré que les techniques de quantification du modèle réduisent considérablement la taille du modèle et la charge de calcul des modèles CNN, permettant le déploiement de modèles CNN sur des dispositifs médicaux portables. Kumar, Abhinav et al. [6] ont proposés un modèle CNN efficace et léger pour la classification d'images histopathologiques (HIC) basé sur MobileNet. Ils ont indiqué qu'en spécifiant un paramètre de profondeur et perfomant une quantification de 16 bits, le modèle proposé est équilibré en terme de précision, temps d'inférence et les exigences de pic de mémoire. Comparé aux modèles pré-entraînés, MobiHisNet a moins de paramètres et de calculs, ce qui permet une classification plus rapide des images.

Dans le but de trouver la méthode de compression la plus efficace qui permet d'avoir une bonne qualité d'image médicale avec une courte période de compression, Bekki et Korti [7] ont proposé trois méthodes de compression : la méthode de la transformée en cosinus discrète (DCT), la transformée en ondelettes discrète (DWT) et méthode de détection compressée(CS). Ils ont trouvé que la méthode DWT donne une meilleure qualité d'image par rapport aux autres méthodes et que la méthode CS est plus rapide que les autres méthodes.

3. CONCEPTS DE BASE

3.1 Compression de modèle

La compression de modèle signifie la réduction de la taille et le temps de réponse par rapport au modèled'origine sans compromettre la précision du modèle. En effet, compréssé un modèlesignifieavoir un modèle avec moins de paramètres et utilisedoncmoins de RAM au moment de l'éxécution et moins de temps pour effectueruneprédiction, ouuneinference.

Il existe plusieurs techniques utilisées pour la compression des réseaux profonds telles que la méthode d'élagage, la quantification et la distillation des connaissances. Dans notre travail nous nous intéressons à la quantifiquation.

3.1.1 Quantification

La quantification vise à accélérer le processus d'exécution en réduisant la complexité de la représentation des nombres et de l'arithmétique et de la logique des operations [8]. Alors que la plupart des standards d'implémentations DNN représentent les pondérations et les activations à l'aide du type de données float32, la quantification permet de représenter ces valeursen utilisant un type de données plus petit (float16 ou int8) [9]. La fonction de quantification est la suivante [10] :

$$r = S(q - Z) \tag{1}$$

Ou, r est la valeur réelle (généralement float32). q est sa représentation quantifiée sous forme d'entier n-bit (uint8, uint32, etc.); S (float32) est le facteur d'échelle, un entier positif qui spécifie la taille du pas de la quantification. Il est utilisé pour mapper la plage dynamique à la plage de format entier $[r_{min}, r_{max}]$ calculé par $[-2^{n-1}, 2^{n-1} - 1]$ [11] :

$$S = \frac{r_{max} - r_{min}}{(2^n - 1)} \quad (2)$$

Z est le "point zéro" quantifié qui reviendra toujours exactement à 0 :

$$Z = round\left(\frac{r_{min}}{S}\right) \quad (3)$$

3.2 Compression d'image

Pendant le développement des classificateurs de réseaux neuronaux, la présence d'un trop grand nombre de caractéristiques d'entréepeut alourdir le processus d'apprentissage et produire un réseau de neurones avec plus de poids de connexion que ceux requis par le problème [12].
De plus, l'inférence est coûteuse en temps de calcul en raison du potentiel élevé dimensionnalité des données d'entrée (par exemple, une haute resolution image) et des millions de calculs qui doivent être sur les données d'entrée. [13]

L'objectif de la compression d'image est de réduire la redondance et la non-pertinence présentes dans l'image, afin qu'elle puisse être stockée et transférée efficacement.

3.2.1. Transformée en ondellette discrète

La compression par ondelettes offre une approche qui permet de réduire la taille des données tout en améliorant leur qualité grâce à la suppression des composantes de bruit à haute fréquence [14].
La transformée en ondelettes discrète (DWT) est une transformée qui décompose un signal donné en un certain nombre d'ensembles, où chaque ensemble est une série temporelle de coefficients décrivant l'évolution temporelle du signal dans la bande de fréquence correspondante [15].

4. APPROCHE PROPOSEE

Dans cette section, La conception d'un système d'aide au diagnostic de la maladie coronarienne est présentée. Ce système vise à optimiser les modèles dee learning permettant leurs utilisationssur les périphériques à ressources limitées en investiguant deux approches de compression : basée modèle et basé images. Ce travail comporte deux phases : phase d'entrainement et phase de test. Le diagramme ci-après représente les principales étapes de la première phase entrainement.

4.1. Phase d'entrainement :

4.1.1. Prétraitement de la base de données

Un des problèmes que nous avons rencontrés dans notre travail est bien le manque en matière de bases de données dans le domaine cardiaque, objet de notre cas d'étude. Il nous a fallu un temps considérable de recherche pour s'approprier une telle BDD. Aussi, pour que cette dernière soit adaptée à notre modèle deeplearning, nous devions revoir sa structure. En effet, les images étaient classées dans des sous-dossiers selon une hiérarchie à niveaux multiples ce qui se traduisait par un chemin d'accès long. Nous avons donc réorganisé la structure de notre BDD de façon à avoir deux dossiers essentiels et rendre ainsi les images facilement accessibles.

4.1.2. Classification

Dans cette étude, nous avons implémenté des modèles CNN en utilisant VGG16. L'architecture VGG16 se compose de cinq blocs avec des couches de convolution et de maxpooling. Les deux premiers blocs sont constitués de deux couches de convolution et les trois derniers blocs se composent de trois couches de convolution suivies d'une couche de max-pooling. ReLU est la fonction d'activation utilisée dans chaque couche de convolution pour prédire la meilleure sortie. Une couche dense en VGG 16 est utilisée pour transférer les neurones des canaux d'entrée vers les canaux de sortie. La couche sigmoïde est utilisée pour la classification.

La phase d'extraction de caractéristiques où Transfer Learning est utilisé pour fournir aux nouveaux réseaux des caractéristiques apprises dans un autre jeu de données. Dans ce sens, on supprime les couches denses du CNN pré-entrainé (vgg16) et on conserve les couches à base convolutivepour qu'on puisse exploiter le modèle sur notre base de données on doit ajouter une couche aplatissement et une couche entièrement connectées personnalisée afin d'effectuer notre classification.

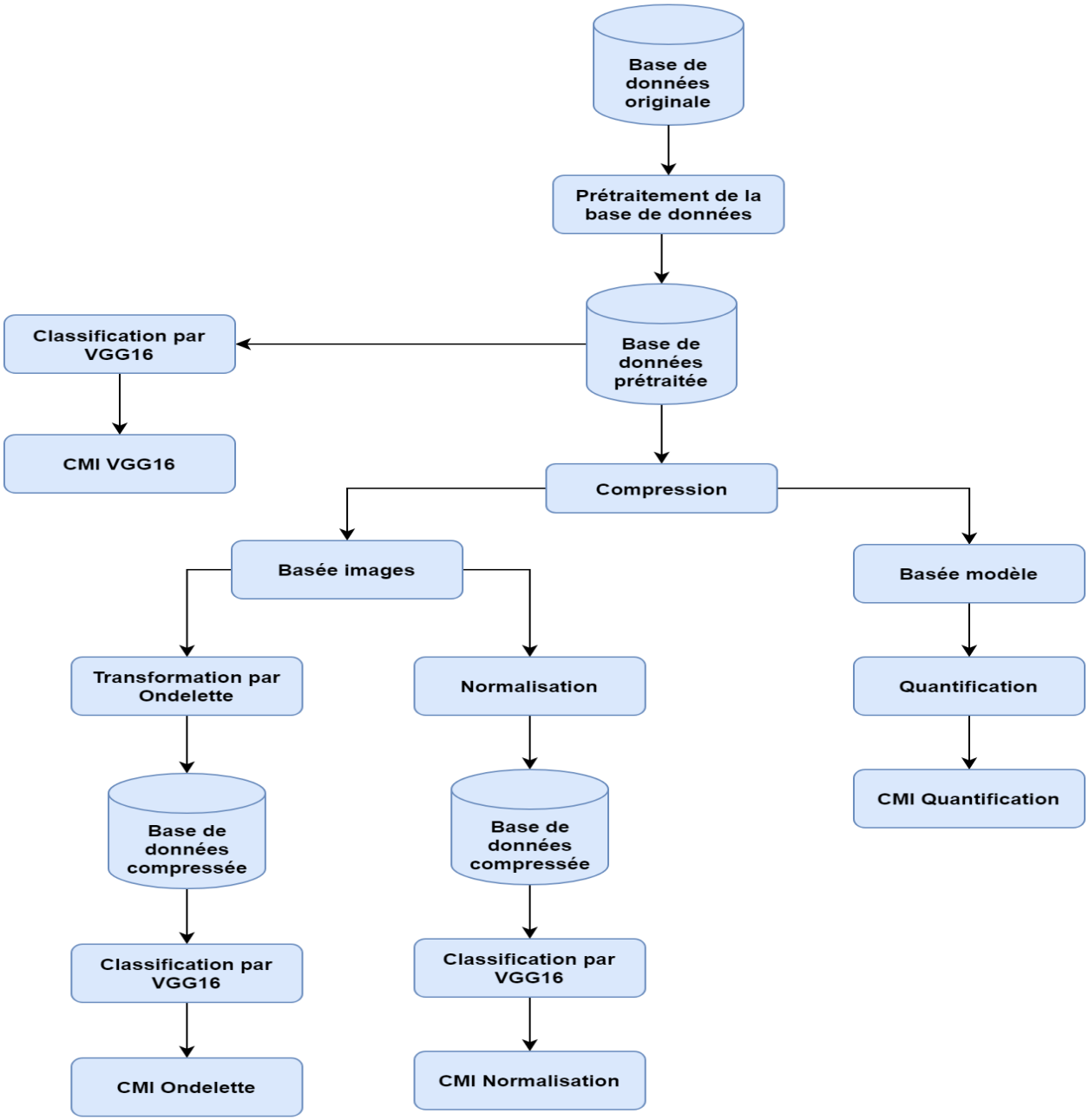


Figure 1 : Phase principale du système proposé

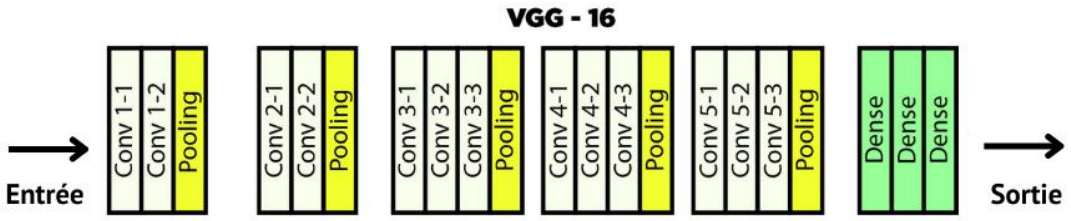


Figure 2: Architecture VGG16

.

4.2. **Quantification**

La quantification peut soit être introduite pendant l'apprentissage ou elle peut être appliqué à un modèle pré-entainé, connu sous le nom de quantification post-formation. Dans ce travail, la technique de post entraînemment est adoptée qui est : *la quantification de plage dynamique* ; cette dernière a pour principe la conversion statique des poids. Lors de l'inférence, de virgule flottante en 8 bits dans la plage symmetrique « [-127, 127] » avec le point zéro seront égal à 0. Cette conversion est effectuée une seule fois et mise en cache pour réduire la latence. Pour améliorer encore la latence, les opérateurs de plage dynamique quantifient dynamiquement les activations en fonction de leur plage à 8bits, avec un point zéro dans la plage asymmetrique « [-128, 127] ». Cependant, les sorties sont toujours stockées en virgule flottante [8].
La quantification des poids se fait statiquement car dans la plupart des cas, les paramètres sont fixés lors de l'inférence. Cependant, les activations sont quantfiés dynamiquement car les cartes d'activation se diffèrent pour chaque échantillon d'entrée [9].

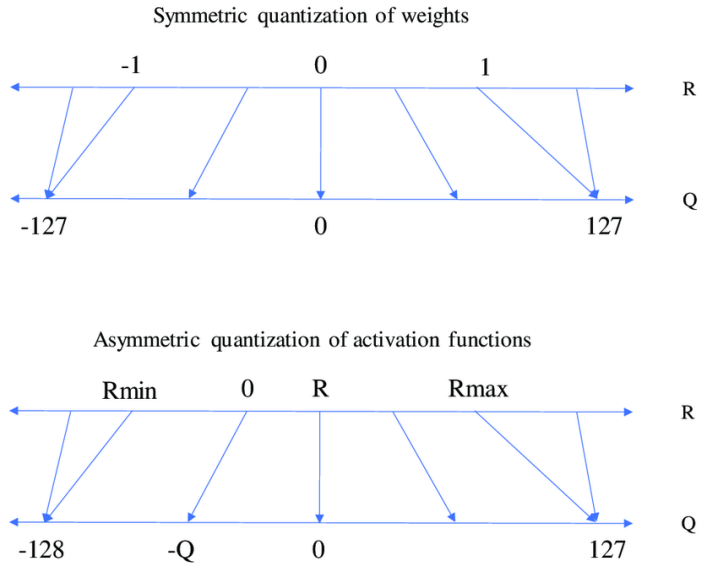


Figure 3 : Quantization des poids et d'activations [8].

4.3. Normalisation

Cetteméthode consiste a exploité les deux fonctions resize () et save () de la bibliothèque Pillow .Nous avons eu recours à la méthode resize(), en passant un argument de tuple à deux entiers représentant la largeur et la hauteur de l'image redimensionnée, elle renvoie à la place une autre image avec les nouvelles dimensions et on spécifie un paramètre de filtre antilalias pour minimiser le crénelage : avoir une apparence de bords plus lisses et une résolution d'image plus élevé . Ensuite on fait appelle à la fonction save () pour enregistrer notre nouvelle image de taille réduite et on spécifie les deux paramètres suivant :

Quality : la qualité de l'image, sur une échelle de 0 (la pire) à 95 (la meilleure), ou la chaîne à conserver. La valeur par défaut est 75. Les valeurs supérieures à 95 doivent être évitées.
Optimise : si présent et vrai, fait compresser la palette en éliminant les couleurs inutilisées.

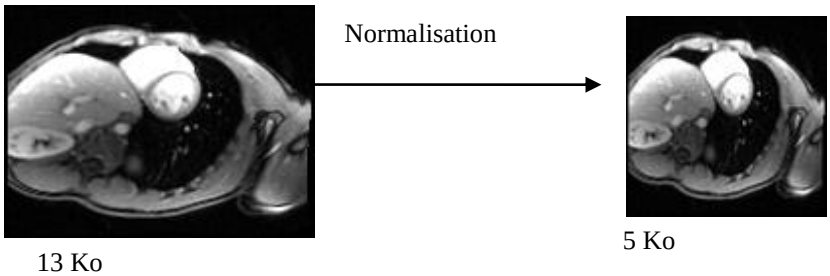


Figure 4: exemple de normalisation

4.4. Transformée en ondellette discrète

Comme Les images sont des signaux 2D, ils sont considérés comme des matrices avec N lignes et M colonnes. A chaque niveau de décomposition, les données horizontales sont filtrées, puis l'approximation et les détails produits à partir de cela sont filtrés sur des colonnes. A chaque niveau, quatre sous-images sont obtenues; l'approximation, le détail vertical, le détail horizontal et le détail diagonal. [16]
La figure ci-dessous présente une décomposition de premier niveau de la DWT :

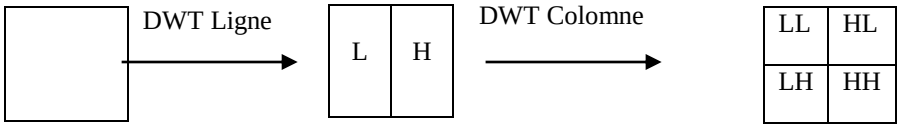


Figure 5 : Décomposition de premier niveau

Nous avons effectué une transformatée en ondelette de premiers niveaux et voici ci-dessous un exemple sur une image de notre base de données :

13 Ko 9 Ko

Figure 6 : Exemple de transformatée en ondelette discrète

4.5. Phase de test

Une fois que l'apprentissage des différents modèles CMI (CoronaryMedical Imaging) de classification de notre base de données est achevé, ces modèles seront sauvegardés pour l'utilisation dans l'analyse des performances en utilisant la base de test ou pour assurer la classe de n'importe quelle image.
Dans notre étude et comme nous nous intéressons à l'optimisation des modèles de classe basée Deep Learning les critères maintenus se concentre essentiellement en plus de la performance du modèle :sa taille, le temps d'inférence ainsi que le nombre de paramètres du modèle générer.
La figure7 illustre le processus de test (chacun des 4 modèles va générer sa décision locale qui peut être normal ou atteint de la maladie coronarienne) :

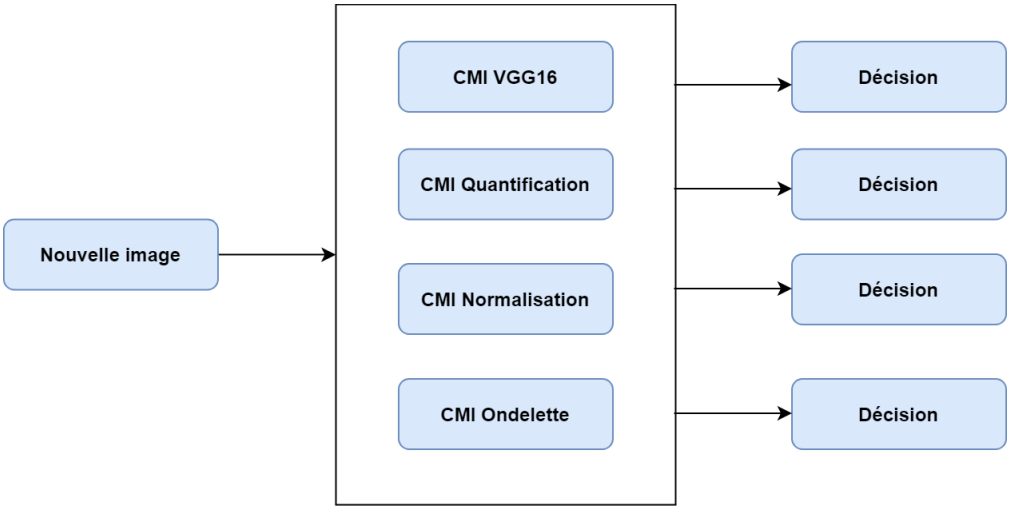
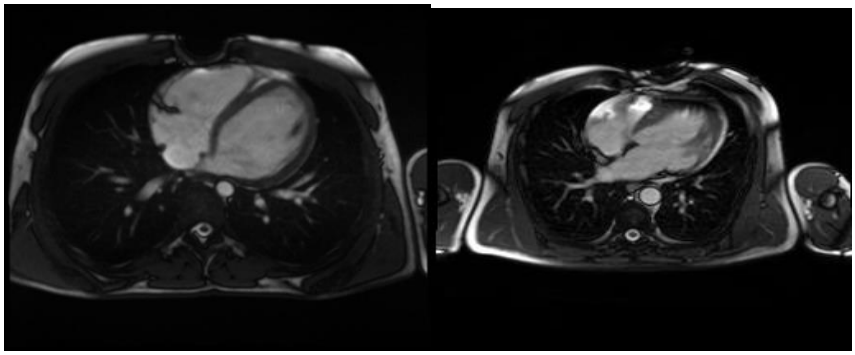


Figure 1: Phase de test

5. Résultats éxpérimentaux

5.1. Base de données utilisée

CAD Cardiac MRI Dataset est une base de données de diagnostic des maladies des artères coronaires basé sur l'imagerie parrésonance magnétique cardiaque. Cette base contient 60000 images appartenant à deux classes : Normal et Malade. Comme notre travail est basé sur le Transfer Learning, L'utilisation d'un nombre très important d'échantillon n'est pas nécessaire. La taille de la base de données utilisée pour la conception de notre modèle contient 13000 images. La base de données est divisée en deux parties : une partie d'entrainement (10 000 images) et une partie pour la phase de test (3000 images). Les données d'apprentissage sont utilisées pour entraîner les modèles tandis que les données de test sont utilisées pour déterminer la performance des modèles tout en analysant l'efficacité des méthodes de compressions. Voici ci-dessous un échantillon de notre base de données :



(Normale) (Malade)

Figure 8 : Un échantillon de la base de données utilisée

5.2. Résultats expérimentaux :

Nous démontrons l'impact de la compression d'images et la compressionde modèle sur sur les mesures de précision, taille de modèle, temps d'inférence ainsi que le nombre de paramètres. Les tables et figures ci-dessous illustrent les principaux résultats obtenus des quatre modèles CMI en leurs appliquant la base de test contenant 3000 images (normale malade).

5.2.1 Résultats du modèle CMI VGG16

Les résultats de la précision de notre premier modèle CMI VGG16 après évaluation et selon différents nombres d'époques testé nous est illustré dans la table ci-dessus.

Tableau 1*: Résutats du modèle CMI VGG16*

Epoques	Précision
1	87%
2	91.8%
20	92.15%
30	96.45%

On illustre ci-après les courbes d'apprentissage et de perte générées par le modèle CMI vgg16

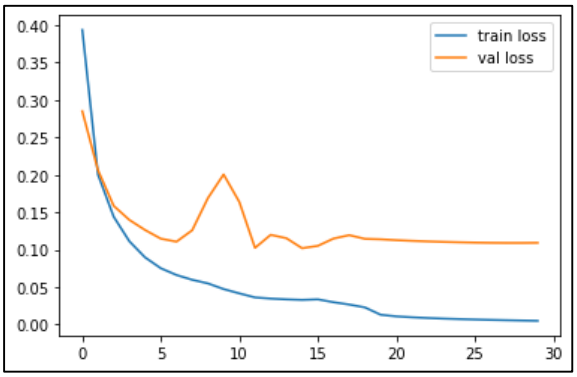


Figure 9 *:* courbe de perte de CMI VGG16

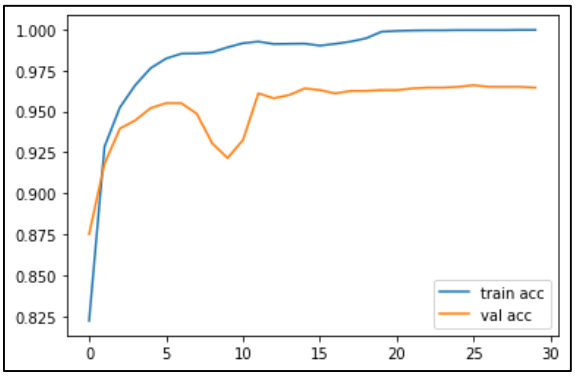


Figure 10 2*:* courbe d'entrainement de CMI *VGG16*

Ce modèle génère les meilleurs résultats avec un nombre d'époques 30. Ces résultats sont satisfaisants.

5.2.2 Résultats du modèle CMI Quantification

La table suivante présente l'impact de la méthode de compression de modèle quantification sur le modèle de base CMI VGG16 en termes de temps d'inférence, taille du modèle et précision.

Tableau 2: Résultats du modèle CMI Quantification

Modèle	Temps d'inférence	Taille du modèle	Précision
CMI VGG16	44.6 ms	56.4 Mo	96.45%
CMI Quantification	38.9 μs	14.1 Mo	96.4%

Nous remarquons que, d'une part, le modèle VGG16 ne subit aucune dégradation de précision lorsqu'il utilise la quantification de plage dynamique. D'autre part, le temps de d'inférence est amélioré de plus de cent fois par rapport a celui obtenu avant compression. Et aussi, la taille du modèle CMI Quantification a diminué de quatre fois la taille du modèle CMI VGG16 après l'application de la plage dynamique.
On conclue qu'en utilisant la quantification, une réduction significative de la taille des modèles et un bon résultat en terme de précision et temps d'inférence sont obtenus, cela permettant le déploiement du complexe modèles CNN aux appareils à ressources limités toutes en gardant la qualité du modèle non quantifié.

5.2.3 Résultats de la compression d'images basée normalisation :

Nos images utilisées comme entrée du modèle CMI VGG16 ont une dimension de 224*224 pixels.Nous avons appliqué à notre base de données initiale trois différentes normalisations aboutissant chacune d'elles à une nouvelle base données.
La table ci-dessous présente les résultats obtenus en matière de nouvelles résolutions d'image, le ratio de compression ainsi que le gain réalisé en espace mémoire. Les deux derniers paramètres sont définis comme suit :

$$\text{Ratio de compression} = \frac{\text{taille d'image sans compression}}{\text{taille d'image compréssée}} \quad (4)$$

$$\text{Gain d'espace} = 1 - \frac{\text{taille d'image compréssée}}{\text{taille d'image sans compression}} \quad (5)$$

Tableau 3: Résultats de la compression d'imagebaséenormalisation

Technique	Résolution d'image	Ratio de compression	Gain d'espace
Normalisation	150*150	1.7 / 1	40%
	100*100	3 / 1	67%
	64*64	5 / 1	80%

5.2.4 Résultats de modèle CMI Normalisation :

La table ci-après présente le nombre de paramètres, la précision et le temps d'inférence en effectuant ces trois compressions pour un modèle CMI Normalisation.

Tableau 4 : Résultat des modèles CMI Normalisation

Ratio de compression	Nombre de paramètres	Précision	Temps d'inférence
5/ 1	2049	91.07%	40.8 ms
3 / 1	4609	93.93%	41.3 ms
1.7 / 1	8193	95.17%	40.3 ms

D'après le tableau 3 et 4, on remarque que :

- D'une part, un ratio de compression élevé produit un gain d'espace mémoire en matière de stockage d'images.
- D'autre part, Sachant que le nombre de paramètres d'un CMI VGG16 de base est 25089, la compression d'images réduit le nombre de paramètre des modèles ce qui implique une réduction d'usage de mémoire RAM lors de l'exécution. Le nombre de paramètres a diminué d'un taux de 92%, 72% et 68% pour les ratios de compression 5/1, 3 /1, 1.7 / 1respectivement.
- En revanche, on remarque que la précision du modèle est inversement proportionnelle au ratio de compression. En d'autres termes, plus ce dernier est élevé moins est la précision du modèle.

Voici ci-après les courbes de perte et de précisons des modèles CMI Normalisation :

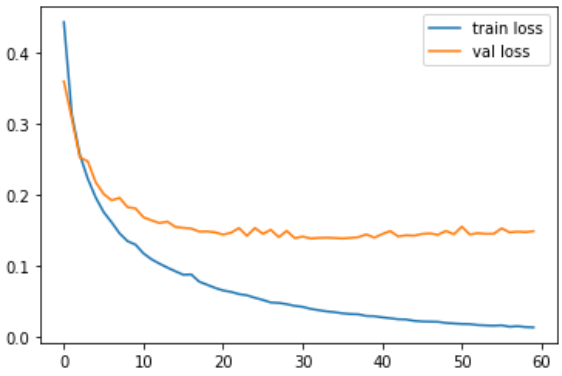


Figure 11 : courbe de perte de CMI Normalisation1.7/1

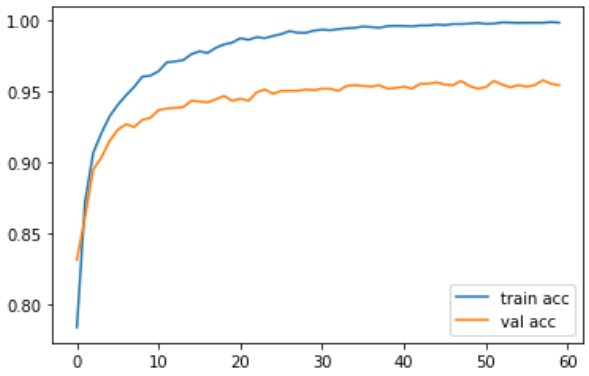


Figure 12 : courbe de précision de CMI Normalisation 1.7/1

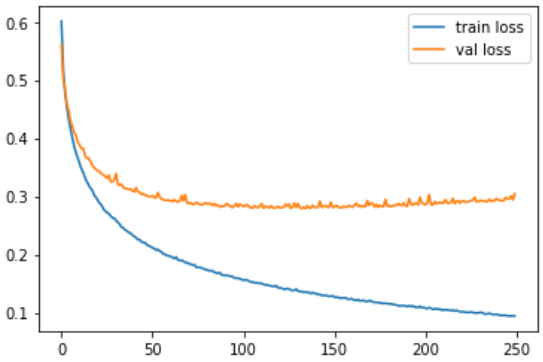


Figure 13: courbe de perte de CMI Normalisation 5/1

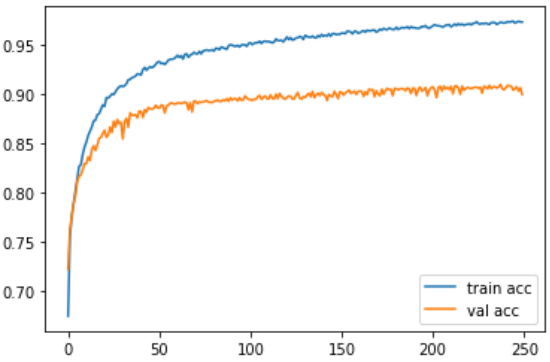


Figure 14:courbe de précision de CMI Normalisation 5/1

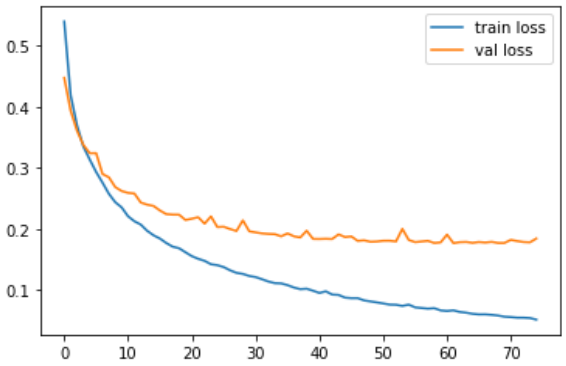


Figure 15: courbe de perte de CMI Normalisation 3/1

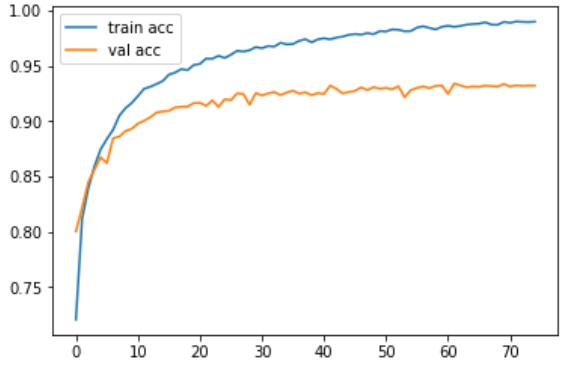


Figure 16:courbe de précision de CMI Normalisation 3/1

5.2.5 Résultats du modèle CMI Ondelette :

Le tableau ci-dessous présente le résultat obtenu après l'exécution du CMI ondelette :

Tableau 5 : Résultat des modèles CMI Ondelette

Modèle	Précision	Temps d'inférence	Taille du modèle
CMI ondelette	96.03%	43.6 ms	56.2 Mo
CMI VGG16	96.45%	44.6 ms	56.4 Mo

Le modèle CMI Ondelette rend une bonne précision par rapport au CMI Normalisation et réduit légèrement le temps d'inférence et la taille du modèle.
Voici ci-dessous les courbes de perte et de précisons des modèles CMI Ondelette :

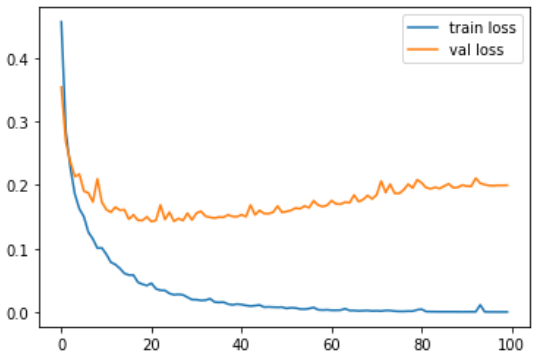


Figure 17 : courbe de perte de CMI Ondelette

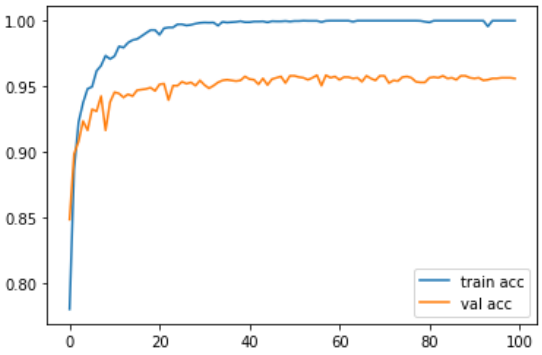


Figure 18 : courbe de précision de CMI Ondelette

6. Conclusion

Le travail présenté analyse un système de classification des images cardiaques par l'apprentissageprofond. Le système adopté est le VGG16 grâce à ces performances dans le domaine médical.
Dans le but d'améliorer la taille du modèle afin de réduire sa complexité et d'éviter l'alourdissement du matériel utilisé des approches de compression ont été proposées.
Nous avons commencé par analyser le comportement de VGG16 sur la base originale et nous avonso btenus des résultatstés satisfaisants.
Pour la phse de compression, nous avons adopté deux modalités basées image et basée modèle.
Pour l'approche basée image, deux techniques ont été investiguées qui sont la normalisation et la transformation en ondelette ; après différents essais sur la valeur de nombres d'époques ; la normalisation a montré qu'il y a une amélioration par rapport au nombre de paramètres du modèle.
Pour l'approche basée modèle, nous avons adopté la méthode de quantification, nous avons conclus qu'il y a un gain ressenti dans la taille du modèle toute en maintenant la précision du modèle.

REFERENCES

[1] Lecun, Y., Bengio, y, & Hinton, g. (2015). Deep learning. *Nature*, *521*(1), 436–444. https://doi.org/10.1038/nature14539.

[2] Touahri, R., Azizi, N., Hammami, N. eddine, Aldwairi, M., Benzebouchi, N. eddine, & Moumene, O. (2021). Multi source retinal fundus image classification using convolution neural networks fusion and Gabor-based texture representation. *International Journal of Computational Vision and Robotics*, *11*(4), 401–428.

[3] Berkaya, S. K., Sivrikoz, I. A., & Gunal, s. (2020). Classification models for SPECT myocardial perfusion imaging. *Computers in Biology and Medicine*, *123*(1), 103893-.

[4] Papandrianos, N., & Elpiniki, P. (2021). Automatic Diagnosis of Coronary Artery Disease in SPECT Myocardial Perfusion Imaging Employing Deep Learning, *Applied Sciences*, *14*(11), 6362-.

[5] Garifulla, M., Shin, J., Kim, Chanho, Kim, H. J., & Hong, S. (2019). A Case Study of Quantizing Convolutional Neural Networks for Fast Disease Diagnosis on Portable Medical Devices. *Sensors*, *22*(1). https://doi.org/10.3390/s22010219

[6] Abhinav, K. (2021). MobiHisNet: A Lightweight CNN in Mobile Edge Computing for Histopathological Image Classification. *IEEE Internet of Things Journal*, *8*(24), 17778–17789.

[7] Bekki, a, & Korfi, A. (2021). Image Processing: Image Compression Using Compressed Sensing, Discrete Cosine Transform and Wavelet Transform. In *Lecture Notes in Networks and Systems: Vol. 413). Artificial Intelligence and Its Applications* (2021st Ed.). Springer.

[8] Shabbeer, S., Basha1, H., Farazuddin1, M., & Pulabaigari, V. (2022). *Deep Model Compression Based on the Training History, eprint arXiv:* (No. 2102.00160). ARXIV.

[9] Stoychev, S., & Gunes, H. (2022). *The effect of model compression on fairness in facial expression recognition* (No. 2201.01709). ARXIV.

[10] Jacob, B., Kligys, S., Kligys, S., Kligys, S., Tang, M., & Adam, H. (2018)., *Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference*. Presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)

[11] MkDoc, A. (n.d.). Neural Network Distiller. Retrieved March 15, 2022, from https://intellabs.github.io/distiller/quantization.html

[12] Batti, R. (2019). Using mutual information for selecting features, supervised neural net learning&quot, *IEEE Transactions on Neural Networks*, *5*(4), 537–550.

[13] Chen, J., & Ran, X. (2019). Deep Learning with Edge Computing: A Review. In *IEEEXplorer*. IEE EXplorer.

[14] Harrington, P. de B. (2016). Multivariate Curve Resolution of Wavelet Compressed Data, *Data Handling in Science and Technology*, *30*, 311–332. https://doi.org/10.1016/B978-0-444-63638-6.00009-7.

[15] Zadeh, M. H. (2021). Robust control applications in biomedical engineering: Control of depth of hypnosis. *Control Applications for Biomedical Engineering Systems*.

[16] Yadav, R., Gangwar, S., & Singh, H. (2012). Study and analysis of wavelet based image compression techniques. *Nternational Journal of Engineering, Science and Technology*, *4*(1).