Classification supervisée d'images de cellules sanguines infectées par la Malaria en utilisant les réseaux de neurones convolutifs

Adnane Bechri: Département de génie des systèmes
Abdallah Maakoul: Département de génie logiciel
École de technologie supérieure
Université du Québec
Montréal, Canada

Mots-clés : Réseaux de neurones convolutifs, classification supervisée, rappel, précision, données d'entraînement.

I. Introduction

La malaria (paludisme) est une maladie qui peut être mortelle, elle est causée par des parasites qui sont transmis par des moustiques infectés de type Anopheles. Cette maladie est présente principalement dans les régions tropicales d'Afrique et d'Asie du Sud-est, mais on peut la retrouver aussi dans de très nombreux pays tels le Mexique, l'Amérique Centrale, l'Amérique du Sud, le Moyen-Orient, etc. La détection de la présence de la maladie dans les cellules sanguines (globules rouges) se fait à travers un examen au microscope et par des techniques de diagnostic sanguin assurées par une expertise humaine. Or la majorité des pays qui sont touchés par la malaria ne possèdent pas les structures médicales pour réaliser ces tests microscopiques. Pour de nombreux malades, on ne peut pas détecter l'infection du paludisme dans les stades très précoces. C'est dans cette optique que s'inscrit le cadre de ce projet, nous visons à introduire un modèle d'apprentissage basé sur les réseaux de neurones convolutifs, permettant de classifier les cellules sanguines en cellules parasitées et non infectées, et ce afin de résoudre le problème de manque de l'expertise dans les pays reculés.

II. Contexte et problématique

La méthode standard la plus utilisée actuellement pour le diagnostic de la Malaria est celle de la microscopie optique des frottis sanguins. Des chercheurs du centre national des communications biomédicales de Lister Hill (LHNCBC), essaient de créer une application mobile capable de classer et de compter les cellules sanguines parasitées, tout ça à partir des images de cellules collectées provenant de 150 patients infectés par la malaria et de 50 patients en bonne santé, et ce dans le but de diagnostiquer rapidement et efficacement le patient dans des zones aux ressources très limitées. Dans ce contexte, plusieurs questions se posent à savoir:

Est-t'il possible d'élaborer un système capable de classifier de telles images en se basant sur une architecture de réseau de neurones convolutif simple?

Avec une telle architecture, peut-on avoir une précision et un taux de rappel assez élevé avec un jeu de donnée d'entraînement de 2000 exemples ?

Lors du déploiement de la solution, peut-on obtenir un système optimal en termes de temps d'exécution?

III. Objectifs

L'objectif de ce projet est de fournir un modèle de classification basé sur les réseaux de neurones convolutifs permettant de :

Classifier les images des cellules sanguines collectées en deux catégories : cellules infectées et cellules saines.

- Quel est l'avantage de l'utilisation d'un algorithme basé sur les réseaux de neurones par rapport à d'autres algorithmes ?
- Quelle est la meilleure architecture de réseaux de neurones à implémenter dans ce modèle ?

Augmenter la performance de ce modèle de classification en optimisant la précision (accuracy), le rappel et le taux d'exécution.

- Comment optimiser les hyperparamètres tels le taux d'apprentissages (Learning rate), le nombre de couches et le nombre d'epochs pour arriver à un modèle performant ?
- Quels sont les meilleurs indicateurs permettant de mesurer la performance de notre modèle ?

Faire en sorte que notre modèle utilise une architecture simple et peu d'exemples en vue d'avoir les meilleures performances possible.

- Peut-on jouer sur le nombre d'epochs, le nombre d'exemples et le nombre de couches d'une manière à augmenter la performance du modèle ?

Optimiser l'utilisation des ressources matérielles (hardware) pour le déploiement du modèle de classification.

- Comment utiliser peu de ressources matérielles de telle sorte à avoir un modèle efficace lors de l'entraînement et de déploiement du modèle ?

IV. Matériels

A. Données :

Les données que nous avons à disposition sont les images des cellules sanguines qui ont été collectées et photographiées à

l'hôpital du Chittagong Medical College au Bangladesh. Ces images ont été étiquetées par un expert de l'unité de recherche "Mahidol-Oxford Tropical Medicine" de Bangkok, en Thaïlande. Au total, 28 558 images de cellules sanguines étiquetées en deux classes égales à savoir : 14 279 cellules saines et 14 279 cellules parasitées. Dans notre cas, l'ensemble des données que nous avons utilisé pour le modèle de classification contient 2 000 exemples divisés également en deux : 1 000 images de cellules labélisées « Uninfected » et 1 000 images de cellules labélisées « parasitized ». Voici un échantillon de deux images étiquetées respectivement "Uninfected" et "Paratized" : **[1]**

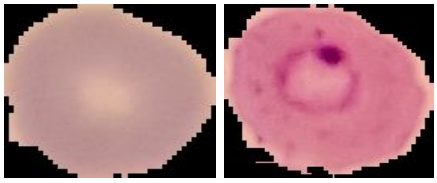

Figure 1: Images de cellule non parasitée/infectée

Le format des images utilisées est le "Portable Network Graphics" (**PNG**). Les dimensions de ces dernières sont fixées à 128 pixels en hauteur et 128 pixels en largeur.

B. Outils :

Les outils que nous avons utilisés pour arriver à nos objectifs sont les suivants :

Python : Ce langage de programmation est la référence en matière de développement des modèles d'apprentissage machine et profond. Il est soutenu par une grande communauté à travers le monde, c'est la raison pour laquelle nous l'avons choisi.

Numpy : Cette bibliothèque nous a permis d'effectuer tous les calculs mathématiques nécessaires lors de la propagation forward et backward et lors de l'optimisation du modèle.

Scipy : Cette librairie a été utilisée pour numériser les images.

Scikit-klearn : Nous avons utilisé la fonction Train_Test_split du module "Cross_validation" pour partager les données entre données d'entraînement et données de test.

TensorFlow : Quant à la partie modélisation, nous avons opté pour TensorFlow qui reste la librairie d'apprentissage machine la plus connue et la plus utilisée. La puissance de TensorFlow en comparaison avec d'autres librairies d'apprentissage machine réside dans le fait qu'on peut créer n'importe quelle application avec, peu importe le degré de difficulté et de la complexité de l'architecture de réseau de neurones utilisée.

V. MÉTHODES

A. Pré-traitement :

Pour la collecte de données, nous avons téléchargé directement l'ensemble des données à partir d'une source proposée par la bibliothèque nationale de médecine (NLM), les images de cellules à classifier ont été déjà étiquetées par un expert. Pour la sélection de données, nous avons créé une fonction que nous avons appelée **«retrieve_names »,** qui permet d'extraire seulement les noms de chacune des images dans les deux dossiers. Une fois appliquée sur le fichier qui contient les données, cette fonction retourne deux listes nommées respectivement: "names" et "labels". La première liste contient seulement les noms des images, la deuxième contient les étiquettes correspondant à chacune des images. Une deuxième fonction que nous avons nommée **« retrieve_data »** reçoit comme argument la liste des noms des images et retourne une liste "images_data" contenant les images numérisées. Pour mélanger ces données, nous avons créé une troisième fonction appelée **« shuffle_data »** qui permute tous les couples image-label pour avoir en sortie un dataset dont les exemples ont été mélangés aléatoirement.

B. Méthodes de forage :

Le modèle d'apprentissage utilisé est basé sur l'architecture de réseau de neurones ci-dessous (figure 2). Vu la simplicité de la tâche de classification des cellules sanguines et qu'il n'y a pas de caractéristiques complexes à identifier dans l'image par le réseau de neurones convolutifs, nous avons décidé de créer une telle architecture simple constituée de 4 couches. Les hyperparamètres ont été choisis de telle sorte que le modèle répond aux objectifs fixés au départ. Voici une liste non exhaustive des hyperparamètres que nous avons choisis avec soin après plusieurs tests d'apprentissage du modèle:

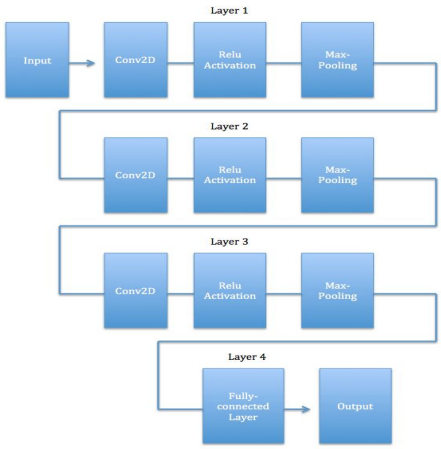

Figure 2: Architecture de réseau de neurones utilisée

Couche 1 : Conv2D : Dimensions des filtres =8, Nombre de filtres = 8, Stride = 1, Padding = same. Relu : Fonction Relu, Relu(z) = Max(0,z). Max-pooling : Dimension de la fenêtre = 8, Stride = 8.

Couche 2 : Conv2D : Dimensions des filtres =4, Nombre de filtres = 16, Stride = 1, Padding = same. Relu : Fonction Relu, Relu(z) = Max(0,z). Max-pooling : Dimension de la fenêtre = 4, Stride = 4.

Couche 3 : Conv2D : Dimensions des filtres =2, Nombre de filtres = 32, Stride = 1, Padding = same. Relu : Fonction Relu,

Relu(z) = Max(0,z). Max-pooling : Dimension de la fenêtre = 2, Stride = 2.

Couche 4 : Fully connected layer.

Learning rate : 0.05 Pour le taux d'apprentissage.

Number of epochs : le modèle a été entrainé 20 fois sur le même dataset.

Mini-Batch size : 64, Nous avons entraîné notre modèle par morceau de 64 exemples pour chaque forward et backward

C. Post-traitement :

Pour ce qui est du post-traitement, nous avons décidé de superviser 3 mesures à savoir le rappel, la précision et l'erreur d'apprentissage. La première mesure est la plus importante, elle quantifie le taux de personnes malades et qui ont été reconnues malades par le modèle par rapport au nombre total de personnes malade, il s'agit bien d'une mesure déterminante dans le domaine de la santé.

Pour le rappel et la précision, nous les avons mesurés sur les données d'entraînement ainsi que sur les données de test, ces mesures sur les données d'entraînement nous ont permis de savoir comment le modèle classifie les images, là où nous lui avons déjà montré leurs étiquettes. Pour visualiser ces mesures, nous avons utilisé la librairie de visualisation des données Matplotlib. Nous avons tracé les quatre graphes correspondant au rappel et la précision sur les données d'entraînement et les données de test en fonction du nombre d'epochs, et ce sur la même figure. Ainsi, nous avons pu comparer les graphes et visualiser l'évolution des 4 mesures en fonction du nombre de fois où le modèle a été entraîné sur les mêmes données. Nous avons également tracé le graphe de l'erreur d'apprentissage en fonction du nombre d'epochs, comme ça, nous avons pu avoir une vision globale sur l'évolution des performances du modèle d'apprentissage en fonction du nombre dans le but de le valider.

VI. RÉSULTATS :

Comme le montre les figures ci-dessous (figure 3 et 4), à l'epoch 20, nous avons :

- Le modèle atteint 100% de précision sur les données d'entraînement (Train accuracy).
- La précision du modèle sur les données de test est de 92 % (Test accuracy).
- Le rappel tend vers 92% sur les données d'entraînement après quelques oscillations (Train recall).
- Le rappel atteint déjà les 92 % sur les données de test
- L'erreur de classification s'approche de 0.

Donc, nous pourrions dire que le modèle atteint déjà des résultats satisfaisants à partir de l'epoch 20, autrement dit, nous pourrions arrêter l'apprentissage à l'epoch 20, la chose qui pourrait nous gagner beaucoup en termes de temps d'apprentissage.

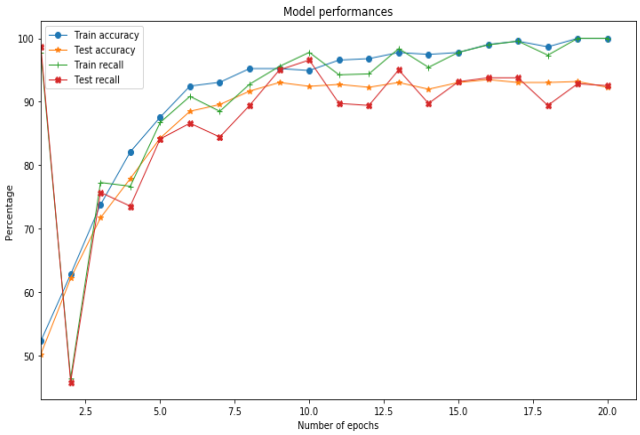

Figure 3: La performance du modèle de classification

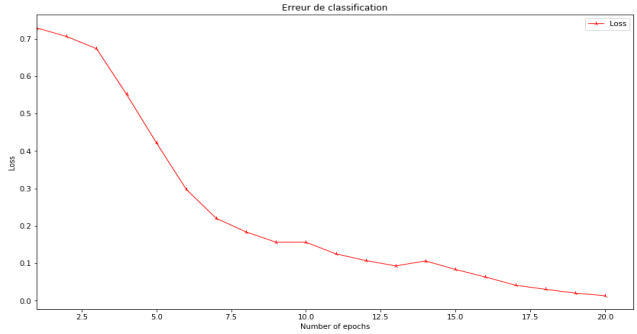

Figure 4: Evolution de l'erreur de classification en fonction du nombre d'epochs

De plus, le modèle risque d'avoir un sur-apprentissage si nous continuons à l'entraîner au-delà d'une vingtaine d'epochs, cela aura comme conséquence le fait que le modèle aurait des difficultés à généraliser les choses apprises lors de l'entraînement, et ce sur de nouveaux exemples.

VII. DISCUSSION :

Dans le cadre de cette discussion, nous avons sélectionné 3 articles qui ont un lien direct avec ce que nous avons fait dans le cadre de cette étude. Ces articles sont :

1. "Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images"

Dans ledit article, les auteurs ont évalué les performances des modèles AlexNet, VGG-16, Xception, ResNet-50 et DenseNet-121, en plus de ça, ils ont créé un nouveau modèle personnalisé de réseau de neurones convolutifs séquentiel. Résultat de cette évaluation : Le ResNet-50 est le meilleur modèle en termes de performances en ayant une précision de 0.957 ± 0.007 et une sensibilité de 0.945 ± 0.020. Quant à nous, nous avons créé un modèle de CNN personnalisé qui a donné 92% de précision et 92% de sensibilité.

2. "Classification of Malaria-Infected Cells Using Deep Convolutional Neural Networks"

Les auteurs de cet article ont décidé d'appliquer un algorithme d'interpolation sur leur dataset, et ce afin d'augmenter les données. Ensuite, ils ont utilisé le modèle LeNet-5 comme modèle d'apprentissage. Nous par contre, nous n'avions pas eu ce besoin d'augmenter les données, car on a un dataset de 28k images, ensuite, nous avons entraîné un modèle que nous avons nous-même conçu, ce dernier a donné des résultats satisfaisants en termes de rappel et de précision.

3. "Malaria parasite detection and cell counting for human and mouse using thin blood smear microscopy".

Dans cet article, les auteurs ont utilisé la méthode d'apprentissage SVM, puis un réseau de neurones artificiel (ANN), après leur apprentissage, le classificateur SVM a atteint une précision de 98% et une sensibilité de 91%, tandis que l'ANN a atteint une précision de 99% et une sensibilité de 90%. Quant à nous, nous avons utilisé uniquement un modèle CNN simple de 4 couches qui a donné une précision de 92% et un rappel de 92%.

Nous pourrions dire que les auteurs des trois articles ont opté pour des méthodes différentes que les nôtres, des réseaux de neurones qui demandent plus ou moins de ressources et des datasets plus ou moins grands, c'est la raison pour laquelle nous avons obtenu des performances différentes, mais qui dépassent toute les 90% en termes de précision et de rappel.

VIII. Conclusion :

Dans le cadre de ce projet, notre objectif était de créer un classificateur de ces cellules sanguines capable de séparer les cellules infectées de la malaria des autres cellules saines. Pour ce faire, nous avons créé un premier programme capable de faire le prétraitement des données fournies par le centre national des communications biomédicales de Lister Hill (LHNCBC). Ensuite, nous avons créé une architecture de réseau de neurones convolutif simple constituée de 4 couches, puis nous l'avons entraîné sur un dataset de 1340 exemples et testé sur 660 exemples, et ce tout en optimisant les hyperparamètres qui ont un impact majeur sur les mesures supervisées. Enfin, nous avons obtenu un modèle avec des performances satisfaisantes permettant de répondre aux objectifs que nous avons fixés au départ à savoir, créer un système de supervision basée sur une architecture simple, et qui donne des performances satisfaisantes en termes de rappel et de précision dépassant les 90 %, tout cela, avec un jeu de données de taille faible.

Pour les chercheurs voulant continuer leur recherche dans le cadre de ce projet, nous recommandons fortement l'optimisation de l'architecture de réseau de neurones présentée dans le cadre de ce projet, et ce afin d'obtenir un rappel qui s'approche de 100% avec moins de ressources, vu le fait qu'il est facile d'extraire les caractéristiques qui distinguent une cellule infectée d'une cellule saine. Nous recommandons également la création d'un autre système capable de compter le nombre de cellules infectées et les cellules saines après les avoir classifiées par ce modèle d'apprentissage, cela aidera le personnel médical à suivre l'évolution de la maladie chez les patients.

Remerciements :

Nous remercions le professeur Sylvie Ratté d'avoir bien voulu corriger le rapport d'étape de ce projet, chose qui nous a permis de bénéficier des précieux conseils et d'alimenter notre réflexion, nous la remercions encore pour sa patience, sa disponibilité et son soutien lors de l'élaboration de ce projet.

Nous remercions les auteurs de l'article ***"Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images"*** *(publié le 16 avril 2018). Cet article était une source précieuse pour la réalisation de ce projet, dont nous avons extrait le dataset des cellules sanguines pour entraîner le modèle de classification.*

Nous remercions également le professeur et le co-fondateur de Coursera Andrew Ng de l'université de Stanford, ses cours sur l'apprentissage profond ont été d'un appui considérable dans ce présent projet.

References

[1] Sivaramakrishnan Rajaraman, Sameer K. Antani, Mahdieh Poostchi, Kamolrat Silamut, Md. A. Hossain, Richard J. Maude, Stefan Jaeger et George R. Thoma, publié le 16 avril 2018, PubMed 29682411, PeerJ, Pre-trained convolutional neural networks as feature extractors toward improved malaria parasite detection in thin blood smear images, https://peerj.com/articles/4568/, consulté le 28 juin 2019.

[2] Stefan Jaeger, Hang Yu, Sameer Antani, Sivaramakrishnan Rajaraman, Feng Yang, Rick Fairhurst, NIH : National Institutes of Health, U.S. National Library of Medicine, Lister Hill National Center for Biomedical Communications, Communications Engineering Branch, Malaria Screener, https://ceb.nlm.nih.gov/projects/malariascreener/?fbclid=IwAR0hfqCIYexuINO16f4nVods7oCm8n-trJUobtTToB4e2-rgjrAzYfqa2Q8 ,consulté le 22 juin 2019.

[3] Coursera: Convolutional Neural Networks, Andrew Ng, https://fr.coursera.org/learn/convolutional-neural-networks?specialization=deep-learning, consulté le 04 juin 2019.

[4] Passeport Santé,Le paludisme (malaria), https://www.passeportsante.net/fr/Maux/Problemes/Fiche.aspx?doc=paludisme_pm&fbclid=IwAR3eLD6krroHpewBD-Qhm6m4Q5lK1xl0bIQKu9J_gpwHaFtOyJlBySaRWr0, consulté le 25 juin 2019.

Source de données : *L'ensemble des données utilisées lors de la présente étude est disponible sur :* https://ceb.nlm.nih.gov/repositories/malaria-datasets/