

7 MAI 2021

EPITECH - ZOIDBERG2.0

RAPPORT D'ETUDES

Zoidberg2.0

GROUP17

NOTRE DÉMARCHE

Ce projet consiste à créer une ou plusieurs intelligences artificielles entraînées par machine learning. Ces intelligences doivent pouvoir analyser des radiographies de poumons et détecter si le patient est atteint d'une pneumonie ou non. Pour les patients atteints de pneumonie, l'intelligence artificielle doit reconnaître le type de pneumonie: virale ou bactérienne.

Pour répondre à ces deux problématiques, nous avons décidé de faire plusieurs intelligences artificielles. Ainsi, nous avons utilisé python3 avec les librairies TensorFlow et PyTorch. Nous avons décidé d'utiliser deux librairies différentes afin de nous aider dans notre choix final d'intelligence artificielle. Afin de réaliser cela, nous avons travaillé en plusieurs étapes.

DÉTECTION DE PNEUMONIE

La première phase fut une longue phase de recherches, de découvertes et d'essais afin d'apprendre à créer des réseaux de neurones servant à faire de la reconnaissance d'images. Durant cette phase, nous nous sommes également renseignés sur le processus de détection de pneumonie que les médecins utilisent. A la fin de cette étape, nous avions réussi à avoir différentes intelligences artificielles (une fonctionnant avec TensorFlow et l'autre avec Pytorch) pouvant indiquer à plus de 90% de précision si un patient est atteint ou non de pneumonie. Nous utilisions un système de classification avec un réseaux de neurones de convolution (CNN).

Ces intelligences étaient entraînées avec les sets de données fournis. Lors de la récupération des images, nous redimensionnons l'image à une taille inférieure (200px par 200px) et nous lui appliquons un grayscaling. Le grayscaling permet de convertir une image en couleur en une nouvelle image qui contiendra uniquement des nuances de gris. Ce traitement de l'image permet d'augmenter les performances du modèle que notre intelligence utilisera.

Importance of grayscaling -

- **Dimension reduction:** For e.g. In RGB images there are three color channels and has three dimensions while grayscaled images are single dimensional.
- **Reduces model complexity:** Consider training neural article on RGB images of 10x10x3 pixel. The input layer will have 300 input nodes. On the other hand, the same neural network will need only 100 input node for grayscaled images.
- **For other algorithms to work:** There are many algorithms that are customized to work only on grayscaled images e.g. Canny edge detection function pre-implemented in OpenCV library works on Grayscale images only.

Explication de l'importance du grayscaling dans le traitement d'images

Source: <https://www.geeksforgeeks.org/python-grayscale-of-images-using-opencv/#:~:text=Grayscale%20is%20the%20process%20of,complete%20black%20and%20complete%20white.>

Toute intelligence artificielle fonctionne avec un modèle. Ce modèle permet d'entrainer l'intelligence à détecter des similarités entre des données. Dans nos modèles, nous utilisions le principe de convolution. La convolution est une opération permettant d'appliquer un filtre à une image pour en extraire certaines caractéristiques.

Une fois le model créé, nous l'avons entraîné avec une dizaine de cycles (dits "epochs"). A la fin de cet entrainement, nous avons testé nos intelligences artificielles sur les sets de validation et nous atteignons une précision de plus de 90%. Nous sommes montés à 98% de précision avec un nombre de cycles supérieurs à vingt lors des phases d'entraînements.

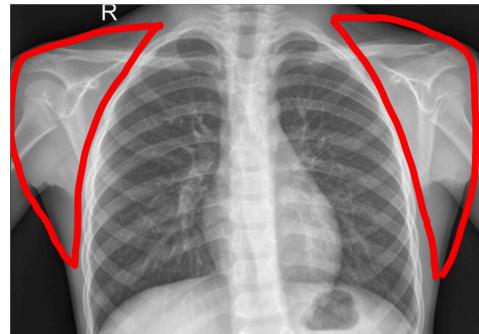
PRÉ-PROCESSING DES DONNÉES

A ce stade, nous avions des intelligences artificielles fonctionnant par convolution avec une précision de plus de 90%. Néanmoins, nous voulions augmenter les performances de nos intelligences en faisant un travail de pré-processing sur les sets de données.

Lors de l'entraînement d'une intelligence artificielle, elle se base sur un set de données. Plus ce set de données est précis et complet, plus le réseau de neurones entraîné sera précis. Lors de la première phase de notre projet, nous avions remarqué des incohérences dans les sets de données. Ainsi, nous avons travaillé sur ces incohérences afin que l'intelligence artificielle s'entraîne avec un set de données le plus cohérent possible.

Afin de "dépolluer" nos sets de données, nous avons exploré plusieurs pistes.

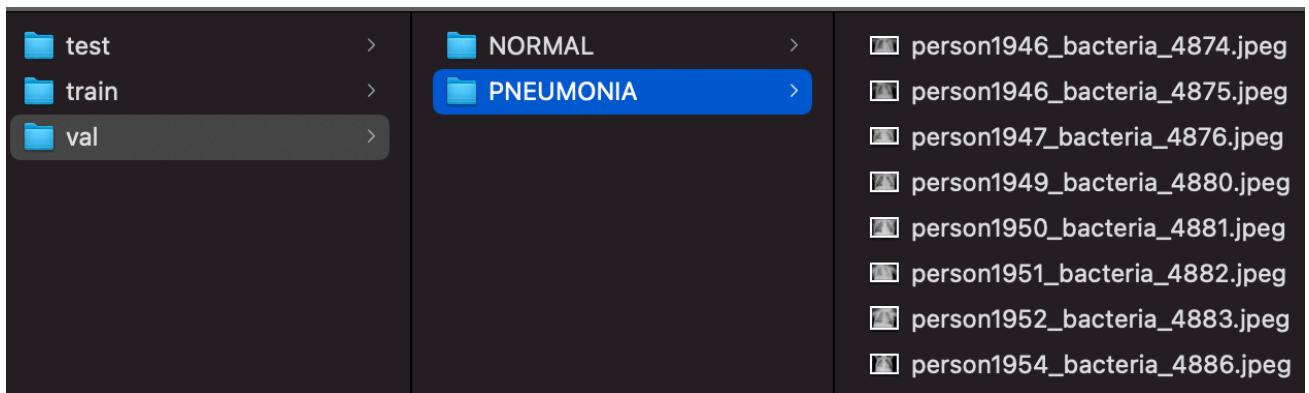
La première piste consistait à garder uniquement les zones de l'image qui nous intéressaient pour détecter la pneumonie. Nous voulions donc enlever toutes les zones de bras et d'épaules des images afin de garder uniquement la cage thoracique. Pour ces zones (bras et épaules) nous n'avions pas remarqué de différences chez les patients sains et malades. Nous avons donc estimé que les enlever ne nuiraient pas à la baisse la précision de nos réseaux de neurones. Cela permet également à nos modèles de s'entraîner uniquement sur ce qui nous intéresse: les variations des poumons dans la cage thoracique.



A gauche: image initiale du set de données d'entraînement.

A droite: les zones entourées en rouge sont les zones que nous voulions enlever de l'image.

La deuxième piste consistait à avoir le même nombre d'image de pneumonie virus ou bactérie dans le set de données d'entraînement. Imaginons que dans le set de données d'entraînement, 80% des pneumonies sont des pneumonies virales alors le réseau de neurones sera plus entraîné à détecter des pneumonies virales que bactérienne. Nous voulions éviter ce biais et entraîner le plus homogènement nos intelligences dans la détection du type de pneumonie. Nous avons également détecté que dans le set de validation, toutes les pneumonies étaient des pneumonies bactériennes. Nous avons donc rajoutés des pneumonies virales dans le set de données afin de confirmer la précision de nos réseaux de neurones.



Visualisation des images de pneumonie du set de données de validation.

La troisième piste consistait à rajouter des images de pneumonies dans notre set de données d'entraînement. Lors de cette étape, il faut faire attention à ne pas trop entraîner nos intelligences car une intelligence trop entraînée avec le même set de données risque de détecter peu d'images ne figurant pas dans son set de données d'entraînement.

Ce travail de pré-processing d'images nous a pris beaucoup de temps et les résultats n'ont pas été aussi importants que voulus. Néanmoins, cela nous a permis de nous concentrer plus en détails dans nos sets de données et de mieux comprendre comment optimiser nos set de données.

DETECTION DU TYPE DE PNEUMONIE

La dernière phase de notre projet consistait à entraîner nos intelligences pour détecter le type de pneumonie.

Afin de réaliser cela, nous avons demandé conseils à des professionnels de la santé et nous avons lu des cours de médecine et nous avons comparés les images de pneumonies de nos sets de données. Tout cela nous a permis de confirmer l'hypothèse qu'on avait: une pneumonie virale se détecte en examinant la diffusion des pixels blancs autour des poumons.

Pour apprendre cela à nos intelligences artificielles, nous lui avons envoyés en données d'entraînements uniquement des pneumonies virales ou bactériennes. Nous utilisions également du grayscaling et nous redimensionnons les images (aux mêmes dimensions que pour l'étape 1). Le modèle ressemble beaucoup à celui de l'étape une. En effet, ici aussi nous utilisons un réseau neuronal convolutif.

Avec 10 cycles d'entraînement, nous avons une précision de 80% pour la détection du type de pneumonie.

CONCLUSION

Au travers de toutes nos recherches et essais, nous avons réalisé deux modèles différents. Le premier permet de détecter si le patient est atteint ou non de pneumonie. Le deuxième lui identifie le type de pneumonie dont le patient est atteint. Nous sommes confiants de la précision que nos modèles ont.

Afin d'améliorer nos réseaux, il faudra travailler sur différents sujets dont le pré-processing des sets de données. Lors de nos échanges avec les responsables de ce projet, nous avons compris à quel point le pré-processing des données est fondamental pour un réseau de neurones efficace. Travailler sur cet axe permettra de grandement améliorer la pertinence de nos réseaux.

De plus, nous étions limités par la puissance de calcul de nos machines. Il serait intéressant de pouvoir entraîner nos intelligences avec plus de cycles afin de gagner en précision.