ReadMe File

To begin with, you will need to have access to the Data shared on Kaggle for the VinBigData competition. The easiest way to use the notebooks would be to run them directly on Kaggle as the different notebooks submitted have been run on Kaggle. However, they can be easily modified for any environment. This ReadMe document will be divided into 5 sections, one for each of the notebooks.

1) **AI_VinBigData_Data_Preprocessing**

In this notebook, the data given for the competition will be preprocessed in order to be used by the next notebooks. If you want to run this notebook outside of Kaggle, you just need to modify the last cell of the Setup Section which set up all the directory for the input and output Datas.

Basically, this notebook will convert and resize with padding all the original DICOM Images into JPEG images of size 512. Then, train.csv document will be updated with the shape of the original images. The position of the bounding boxes will be recomputed according to the new size and the padding added. In addition, a csv file will be created in order to store the shape of the original images in the test set. These shapes will be used after the object detection algorithm as the bounding boxes will need to be relocated to their correct position on the original image.

Finally, text files will be created for all the training images in order to store the annotations. Each row of the file will be as follows:

Label x_mid y_mid w h

Where x_mid and y_mid are the relative coordinates of the bounding box and w and h its relative shape. These label files are used by Yolo model when training and validating.

The users need to be informed that images conversion takes a long time. If one wants to run this notebook, it might time 4 to 7 hours to complete depending on the environment

These processed data are available publicly on Kaggle:

- https://www.kaggle.com/datasets/theolange/ai-vinbigdata

2) **AI_VinBigData_Visualisation**

This notebook will need both the initial dataset provided for the competition and the processed data obtained with the previous notebook. If you want to run this notebook outside of Kaggle, you just need to modify the last cell of the Setup Section which set up all the directory for the input and output Datas.

This notebook will be useful to give more insights and information about the dataset. Different figures are displayed to analyse the classes distribution. The impact of the radiologists is also shown.

The position of the bounding boxes is also studied as well as a fusion algorithm.

3) **AI_VinBigData_Yolov8x**

This notebook requires the processed data obtained after preprocessing as input. If you want to run this notebook outside of Kaggle, you just need to modify the last cell of the Setup Section which set up all the directory for the input and output Datas.

This notebook details the preprocessing method used for the preparation of the data to train a Yolov8x model. Datas are separated into a train and a validation set using the KFold method. By default, fold 4 is selected as the validation fold. You can update this by changing the value of the parameter VAL_FOLD in the Setup section of the notebook.

Once the Datas are ready, a Yolov8x model will be trained over 40 epochs with validation and batches of size 16. After training, inferences will be made on the test set.

The bounding boxes will then be relocated according to the shape of the original test images. The notebook will automatically clean the working directory to keep only relevant figures, the mode best weights and the prediction file.

To obtain the 5 models described in the report, one would need to run this notebook 5 times, changing the parameter VAL_FOLD each time. However, the reader needs to be informed that it might take about 4 hours to train a single model.

All the model's best weights and prediction files are available publicly on Kaggle:

- https://www.kaggle.com/datasets/theolange/ai-vinbigdata

4) **AI_VinBigData_ResNet101**

This notebook requires the processed data obtained after preprocessing as input. If you want to run this notebook outside of Kaggle, you just need to modify the last cell of the Setup Section which set up all the directory for the input and output Datas.

This notebook details the whole pipeline for the training, validating and inference process of a ResNet101 model.

The Data will first be divided into two classes and both classes will be evenly divided into 5 Folds. By default, Fold 4 is used for cross-validation. You can update this by changing the value of the parameter VAL_FOLD in the Setup section of the notebook. Then, Data Loaders are created to train, validate and make inference.

The model is then trained using Cross Entropy loss, Stochastic Gradient Descent to optimise its parameters and a Cosine Annealing Warm Restarts learning rate scheduler for 15 epochs.

The accuracy and performance of the trained model will be shown by displaying different figures.

Finally, inferences will be made on the test set in order to get the probability of each image belonging to both classes.

This notebook has been run on GPU for about 45 minutes.

The train model and prediction file are available publicly on Kaggle:

- https://www.kaggle.com/datasets/theolange/ai-vinbigdata

5) **AI_VinBigData_Model_Ensembling**

This notebook requires all the prediction files obtained by training the Yolov8x model and the ResNet101 classifier. If you want to run this notebook outside of Kaggle, you just need to modify the last cell of the Setup Section which set up all the directory for the input and output Datas.

This notebook will apply the ensembling strategy to the prediction made by the Yolov8x models. Then, the 2-class filter will be used according to the prediction score obtained by the ResNet model.

Only the final prediction file is given with these notebooks. The prediction obtained after the ensembling and the 2-class filter are both available on Kaggle:

- https://www.kaggle.com/datasets/theolange/ai-vinbigdata