Model Card and App Info v1.0

Revision Record

App name: XarpAi Lung Opacity Detector
Model name: exp114_model.pt

App deployment status: Prototype
Version: 1.0
Date: 1-Apr-2023
Created by: vbookshelf
Notes: Released for demonstration

Known Issues

1- The latency (prediction time) is around 4 seconds per image. This is because the CPU is being used for inference. Inference would be faster with a GPU.

2- The app sometimes outputs multiple overlapping bounding boxes. I haven't fixed this because this behaviour could have some significance that I don't understand, but a radiologist might. The bounding boxes can be hidden by simply clicking on the image.

3- The app predicts false positives. The app does not detect opacities with 100% accuracy.

-x-

Summary

The XarpAi Lung Opacity Detector is a proof-of-concept for a free, open-source desktop app that uses artificial intelligence to detect opacities on chest x-rays.

Opacities are characteristic signs of lung diseases like TB and Pneumonia. This app analyses chest x-rays and draws bounding boxes around opacities. Radiologists can then review these areas of interest and use their clinical judgement to make a final diagnosis.

There's a shortage of radiologists in poor countries. In 2015, Rwanda had 11 radiologists to serve its population of 12 million people. Liberia, with a population of four million, had two practising radiologists. This app provides high volume diagnosis support. It can help overwhelmed radiologists triage x-rays and speed up their workflows.

The predictions are made by a Pytorch Faster R-CNN model. The model was fine tuned on data from four chest x-ray datasets:

- The TBX11K Tuberculosis dataset
- The Kaggle RSNA Pneumonia Detection Challenge
- The Kaggle VinBigData Chest X-ray Abnormalities Detection competition
- The Kaggle SIIM-FISABIO-RSNA COVID-19 Detection competition

Although the app displays opacity bounding boxes, the model is also trained to detect lungs i.e. it predicts a bounding box that surrounds both lungs. If the model fails to detect the lungs then the app outputs an error message.

The model was validated on an 80/20 train test split. It was also tested on three out of sample datasets:

- The Shenzhen and Montgomery Tuberculosis datasets
- The DA and DB Tuberculosis datasets
- Child Chest X-Ray Pneumonia dataset

These out of sample datasets don't have annotated opacity bounding boxes. Therefore, accuracy was used as a rough metric - if the target was positive (e.g. positive for TB) and the model predicted a bounding box, the model was deemed to have made a correct prediction. This validation approach is not rigorous. But it's a quick and simple way to get a feel for the model's capability.

Results on the 20% validation data:
map@0.5: 0.776
accuracy: 0.91

Accuracy on out of sample datasets:
- Shenzhen and Montgomery TB datasets: 0.84
- DA and DB TB datasets: 0.85
- Child Chest X-Ray Pneumonia dataset: 0.83

Chest x-rays are difficult for humans to read. One study (TBX11k paper) found that radiologists have a 68.7% accuracy when diagnosing TB on chest x-rays. Using that number for context, this model's test results look very good. The good performance on the child pneumonia data is surprising because the training data didn't include a large number of child x-rays.

These results show that this opacity detection app could be helpful when diagnosing lung diseases like TB and Pneumonia. Thank you for trying this app. Please feel free to share your feedback in the discussion forum: https://www.kaggle.com/datasets/vbookshelf/xarpai-lung-opacity-detector/discussion

Patient Data Security

This is a desktop application.
- Patient data never leaves the user’s pc or laptop.
- There's no tracking.
- The code is fully accessible and therefore auditable for malware.

Input

The app accepts chest x-rays in dicom, png and jpg formats. Multiple files can be submitted at the same time.

Output

The app outputs images with bounding boxes drawn around the detected opacities.

Data Summary

Total images: 17,283
Total opacity images: 10,983
Total no_opacity images: 6,300

This data was sampled from four chest x-ray datasets:

- The TBX11K Tuberculosis dataset
    - Total images: 2,299
    - Total opacity images: 799
    - Total no_opacity images: 1,100

- The Kaggle RSNA Pneumonia Detection Challenge
    - Total images: 14,861
    - Total opacity images: 6,011
    - Total no_opacity images: 2,700

- The Kaggle VinBigData Chest X-ray Abnormalities Detection competition
    - Total images: 2,752
    - Total opacity images: 1,322
    - Total no_opacity images: 1,100

- The Kaggle SIIM-FISABIO-RSNA COVID-19 Detection competition
    - Total images: 4,587
    - Total opacity images: 2,851
    - Total no_opacity images: 1,400

Data Pre-processing

Data pre-processing involved converting all images to grayscale, padding the images so that they were square and normalising. Faster R-CNN does not need all images to be converted to the same size.

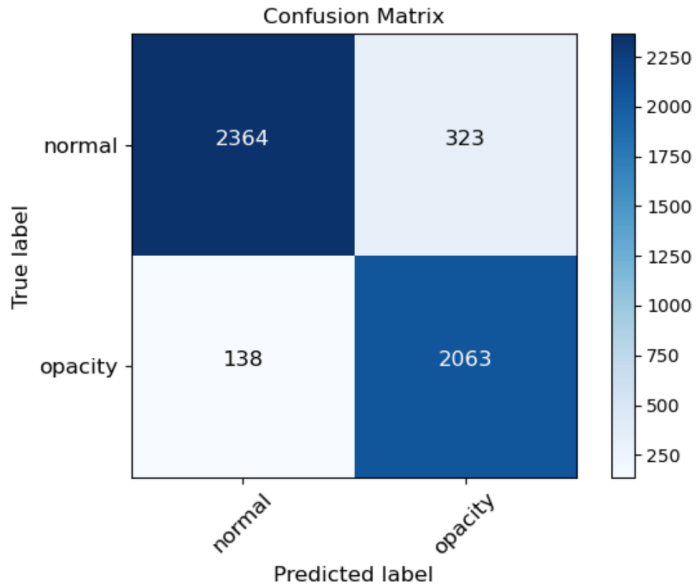
Creating the train and val datasets

The local validation data was created using an 80/20 split. The data was stratified by data source to ensure that data from all four datasets were included in both the training and validation sets. The data was also stratified by target i.e. opacity or no_opacity.

Validation Results

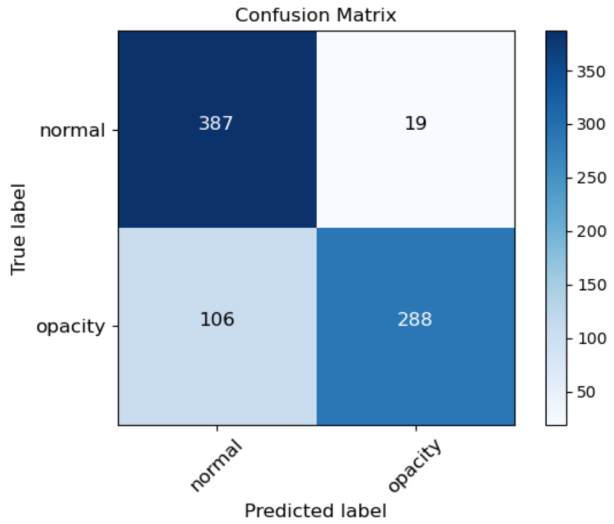
1- Local validation (20% of data)

```
IoU metric: bbox
 Average Precision  (AP) @[ IoU=0.50:0.95 | area=   all | maxD
ets=100 ] = 0.639
 Average Precision  (AP) @[ IoU=0.50      | area=   all | maxD
ets=100 ] = 0.776
 Average Precision  (AP) @[ IoU=0.75      | area=   all | maxD
ets=100 ] = 0.618
```



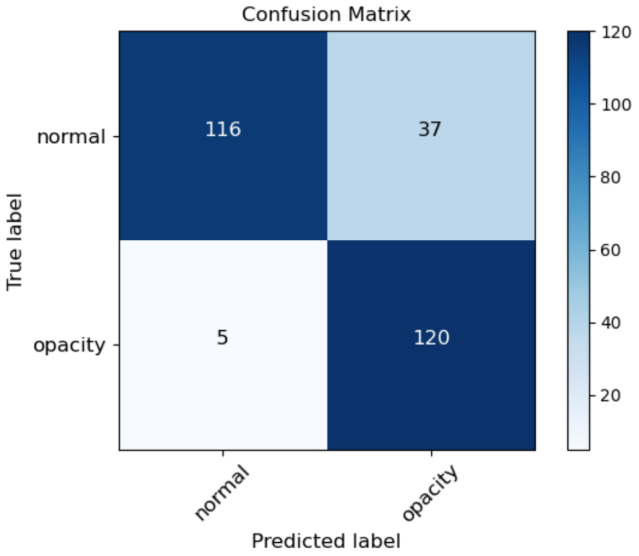
	precision	recall	f1-score	support
normal	0.94	0.88	0.91	2687
opacity	0.86	0.94	0.90	2201
accuracy			0.91	4888
macro avg	0.90	0.91	0.91	4888
weighted avg	0.91	0.91	0.91	4888

2- Shenzhen and Montgomery datasets



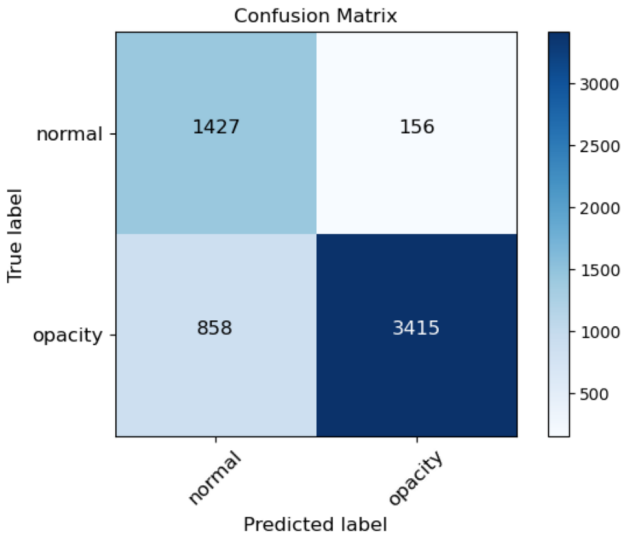
	precision	recall	f1-score	support
normal	0.78	0.95	0.86	406
opacity	0.94	0.73	0.82	394
accuracy			0.84	800
macro avg	0.86	0.84	0.84	800
weighted avg	0.86	0.84	0.84	800

3- DA and DB datasets



	precision	recall	f1-score	support
normal	0.96	0.76	0.85	153
opacity	0.76	0.96	0.85	125
accuracy			0.85	278
macro avg	0.86	0.86	0.85	278
weighted avg	0.87	0.85	0.85	278

4- Child Chest X-Ray Pneumonia dataset



	precision	recall	f1-score	support
normal	0.62	0.90	0.74	1583
opacity	0.96	0.80	0.87	4273
accuracy			0.83	5856
macro avg	0.79	0.85	0.80	5856
weighted avg	0.87	0.83	0.83	5856

Hardware

- 1 x P100 GPU
- Trained on Kaggle
- Training time was approx. 7 hours (8 epochs).

Misc Info

1- In practice users could submit images of varying quality. These images could have been taken using different types of x-ray machines or even be photos of x-ray films taken with cellphone cameras. These and other real-world factors could reduce the accuracy of the model.

2- Although the app produces good test results on paediatric images, there weren't many paediatric images in the training data. In the real world, this app may not produce robust results on paediatric data.

3- The app can easily be configured to use a GPU.

4- I used data augmentation to improve accuracy and reduce overfitting.

Documentation

All code is available on GitHub:
https://github.com/vbookshelf/XarpAi-Lung-Opacity-Detector

Licence

The app code is available under an MIT License. But please note that the trained model can't be used commercially because some of the training data is not licensed for commercial use.

Contact

Email: contact -at- woza -dot- work

-x-

Links to datasets and papers

Shenzhen and Montgomery datasets
Download from Kaggle:
https://www.kaggle.com/datasets/kmader/pulmonary-chest-xray-abnormalities
Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4256233/

TBX11K Simplified
Download from Kaggle:
https://www.kaggle.com/datasets/vbookshelf/tbx11k-simplified
Paper:
https://openaccess.thecvf.com/content_CVPR_2020/papers/Liu_Rethinking_Computer-Aided_Tuberculosis_Diagnosis_CVPR_2020_paper.pdf

DA and DB datasets
Download from Kaggle:
https://www.kaggle.com/datasets/vbookshelf/da-and-db-tb-chest-x-ray-datasets
Download: https://sourceforge.net/projects/tbxpredict/files/data/
Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4229306/

Belarus dataset
Download: https://github.com/frapa/tbcnn/tree/master/belarus
Paper: https://www.nature.com/articles/s41598-019-42557-4

VinDr-CXR dataset
A version of this dataset was used for a Kaggle competition:
https://www.kaggle.com/competitions/vinbigdata-chest-xray-abnormalities-detection/overview
Download: https://physionet.org/content/vindr-cxr/1.0.0/
Paper: https://www.nature.com/articles/s41597-022-01498-w

VinDr-PCXR dataset
Download: https://physionet.org/content/vindr-pcxr/1.0.0/
Paper:
https://www.medrxiv.org/content/10.1101/2022.03.04.22271937v1.full-text

Tuberculosis (TB) Chest X-ray Database
Download from Kaggle:
https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset
Paper: https://ieeexplore.ieee.org/document/9224622

Child Chest X-Ray Images (Version 2)
paultimothymooney dataset version 2 on Kaggle:
https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia
andrewmvd dataset version 2 on Kaggle:
https://www.kaggle.com/datasets/andrewmvd/pediatric-pneumonia-chest-xray
Paper: https://www.cell.com/cell/fulltext/S0092-8674(18)30154-5

RSNA Pneumonia Detection Challenge
Download from Kaggle:
https://www.kaggle.com/competitions/rsna-pneumonia-detection-challenge/data
Download from the RSNA website:
https://www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pneumonia-detection-challenge-2018
Paper: https://pubs.rsna.org/doi/10.1148/ryai.2019180041

SIIM-FISABIO-RSNA COVID-19 Detection
Dowload from Kaggle:
https://www.kaggle.com/competitions/siim-covid19-detection/data
Paper: https://osf.io/532ek/