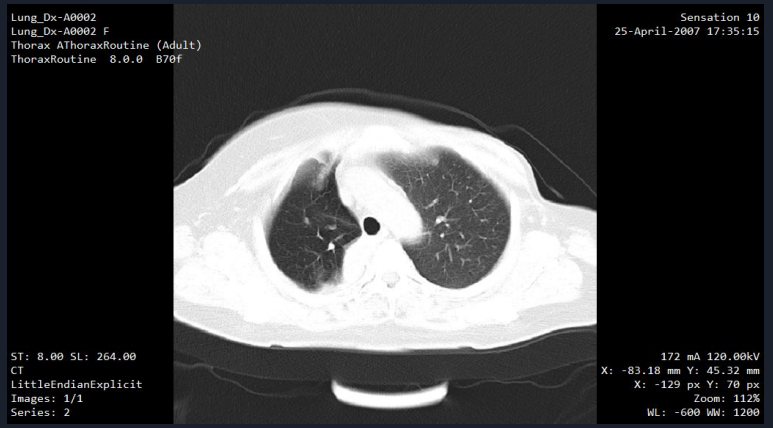
Lung Cancer Image Classification and Bounding Box Predictions

By Daniel Sun

Problem Statement

- Average CT or MRi scans can range anywhere between 4 to 600+ images
- Identifying the location of a cancerous mass can be time consuming to do by hand and requires vast experience and knowledge
- Can I create a model to automate cancer identification and classification?

The Dataset



- Taken from Cancer Imaging Archive: A Large-Scale CT and PET/CT Dataset for Lung Cancer Diagnosis

- Entire dataset:
  - 127gb Dicom files
  - 4 classes of lung cancer
    - Adenocarcinoma, Small Cell Carcinoma, Large Cell Carcinoma, Squamous Cell Carcinoma
  - Hand-annotated bounding boxes in separate XML files

- My subset:
  - ~ 20gb
  - ~1000 labeled images used for training

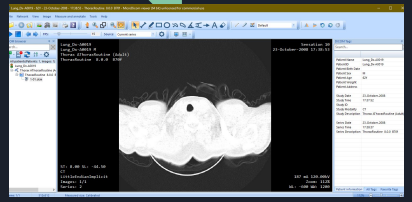
```
<?xml version="1.0"?>
- <annotation verified="no">
    <folder>Unknow</folder>
    <filename>Unknow</filename>
    <path>Unknow</path>
  - <source>
      Unknow
      <database>Unknown</database>
    </source>
  - <size>
      <width>512</width>
      <height>512</height>
      <depth>1</depth>
    </size>
    <segmented>0</segmented>
  - <object>
      <name>A</name>
      <pose>Unspecified</pose>
      <truncated>0</truncated>
      <Difficult>0</Difficult>
    - <bndbox>
        <xmin>286</xmin>
        <ymin>310</ymin>
        <xmax>355</xmax>
        <ymax>402</ymax>
      </bndbox>
    </object>
</annotation>
```

Goals

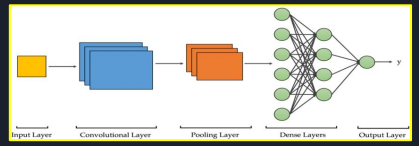
1. Accurately draw a bounding box around a cancerous mass within a given image

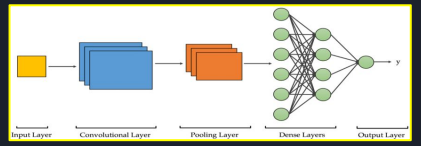
2. Accurately predict the class of cancer found within given bounding box

General Workflow



Training





Evaluate models

Final models

Pull out bounding box predictions

The metric IOU (Intersection Over Union)

- IOU = Area of Overlap / Area of Union：

Model Statistics on Validation Set

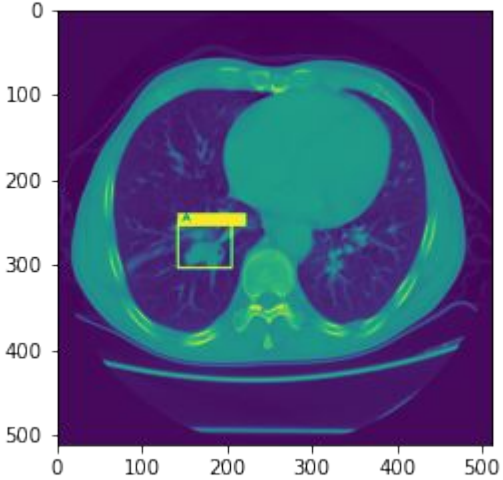
- Ratio of predicted Boxes *with* overlap: 0.772
  - Total testing length: 232
  - Total non-zero IOU: 179

- Average overlap of non-zero IOUS: 0.377

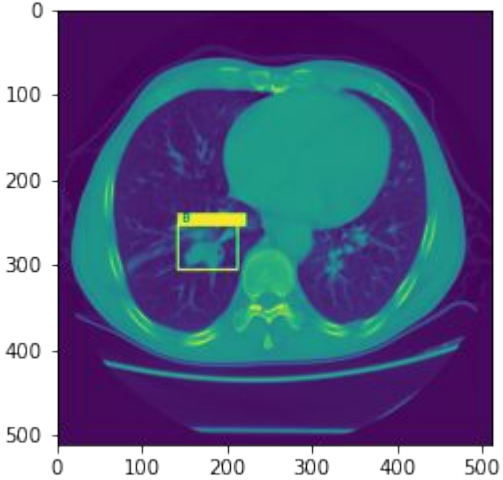
Best prediction

IOU score: 0.870

Model Predicted



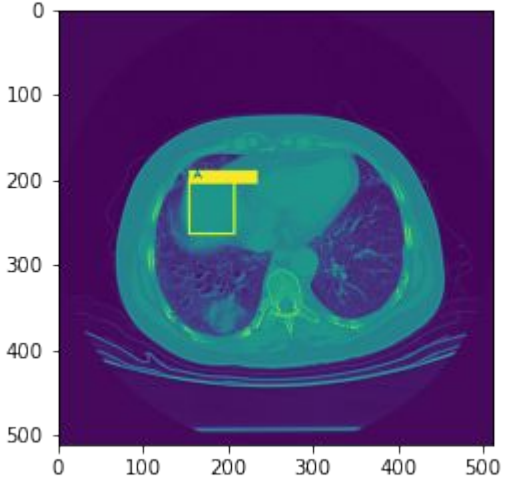
Actual



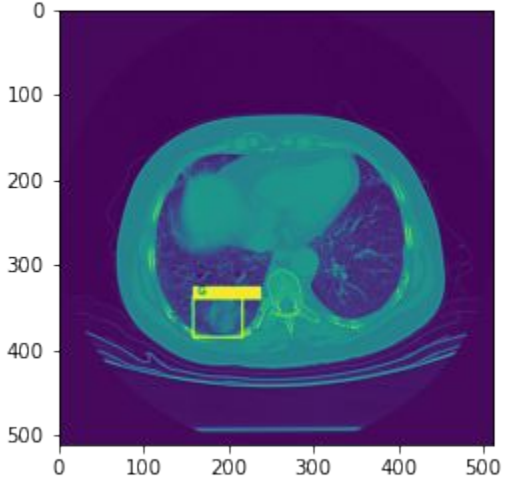
Worst Prediction

IOU score: 0

Model Predicted



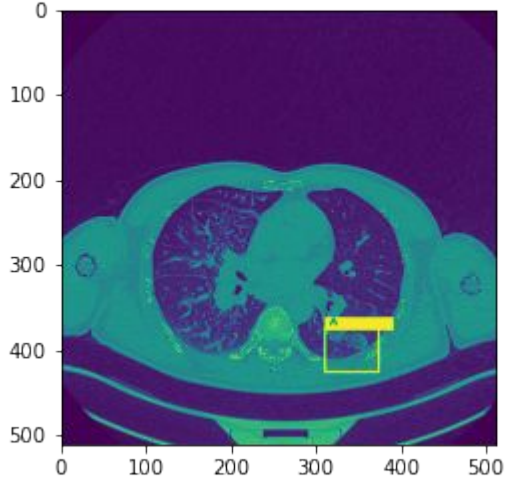
Actual



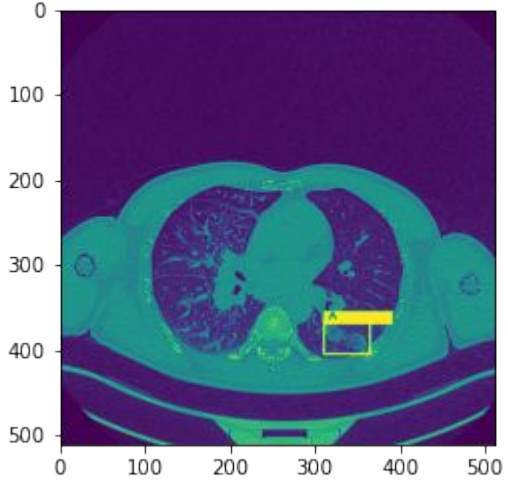
The Average Prediction

IOU score: 0.455

Model Predicted

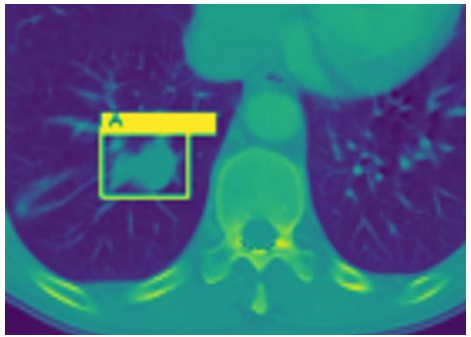


Actual



Classification using Bounding Box versus Image Data

Bounding Box Coord → Predicts: A



Actually: B

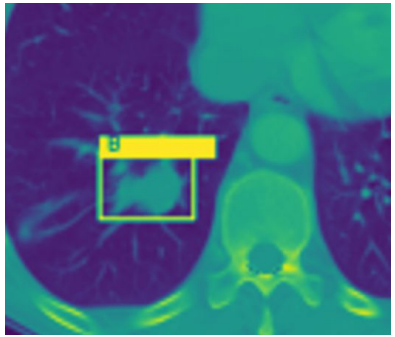
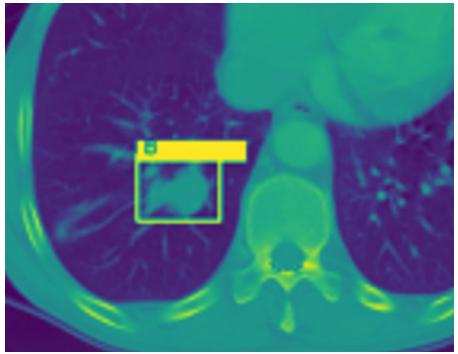


Image Data → Predicts: B



Classification Accuracy

Accuracy using Bounding Box Coordinates to classify			
A	B	E	G
0.830	0.667	0.513	0.671
Weighted F1 Score		0.704	
Accuracy using Image Data to classify			
0.959	0.959	0.889	0.940
Weighted F1 Score		0.947	

Testing on a new set of images

Accuracy using Bounding Box Coordinates to classify			
A	B	E	G
0.349	0.139	0.386	0.468
Weighted F1 Score		0.402	
Accuracy using Image Data to classify			
0.583	0.421	0.107	0.431
Weighted F1 Score		0.452	

- Ratio of predicted Boxes *with* overlap: 0.215
  - Total testing length: 1861
  - Total non-zero IOU: 401
- Average overlap of non-zero IOUS: 0.214

Conclusion

- Model is not ready for production: Accuracy and IOUs need to be much more accurate on testing sets

- Model is niche: trained on images where it is given that there *is* a cancerous mass
  - May be not able to say ‘there is no cancerous mass’ within an image

- Proof of Concept

Future Steps

- Optimize system to handle more complex models and more hypertuning
  - Perhaps utilize cloud gpus for offloading memory pressure
- Begin with a pre-trained neural net like EfficientNet or ResNet50V2
- Utilize a larger training set or the entirety if possible