TensorFlow Lite now supports converting weights to 16-bit floating point values during model conversion from TensorFlow to TensorFlow Lite's flat buffer format. This results in a 2x reduction in model size. Some hardware, like GPUs, can compute natively in this reduced precision arithmetic, realizing a speedup over traditional floating point execution. The Tensorflow Lite GPU delegate can be configured to run in this way. However, a model converted to float16 weights can still run on the CPU without additional modification: the float16 weights are upsampled to float32 prior to the first inference. This permits a significant reduction in model size in exchange for a minimal impacts to latency and accuracy.

Refer - https://www.tensorflow.org/lite/performance/post_training_float16_quant

In [1]:

```
from google.colab import drive
drive.mount('/gdrive')
%cd /gdrive/
```

```
Mounted at /gdrive
/gdrive
```

In [2]:

```
!pip install keras==2.3.1
!pip install tensorflow_io
```

```
Collecting keras==2.3.1
  Downloading https://files.pythonhosted.org/packages/ad/fd/6bfe87920d7f4fd4
75acd28500a42482b6b84479832bdc0fe9e589a60ceb/Keras-2.3.1-py2.py3-none-any.wh
l (https://files.pythonhosted.org/packages/ad/fd/6bfe87920d7f4fd475acd28500a
42482b6b84479832bdc0fe9e589a60ceb/Keras-2.3.1-py2.py3-none-any.whl) (377kB)
     |████████████████████████████████| 378kB 5.8MB/s
Requirement already satisfied: h5py in /usr/local/lib/python3.7/dist-package
s (from keras==2.3.1) (2.10.0)
Requirement already satisfied: six>=1.9.0 in /usr/local/lib/python3.7/dist-p
ackages (from keras==2.3.1) (1.15.0)
Requirement already satisfied: numpy>=1.9.1 in /usr/local/lib/python3.7/dist
-packages (from keras==2.3.1) (1.19.5)
Requirement already satisfied: keras-preprocessing>=1.0.5 in /usr/local/lib/
python3.7/dist-packages (from keras==2.3.1) (1.1.2)
Collecting keras-applications>=1.0.6
  Downloading https://files.pythonhosted.org/packages/71/e3/19762fdfc62877ae
9102edf6342d71b28fbfd9dea3d2f96a882ce099b03f/Keras_Applications-1.0.8-py3-no
ne-any.whl (https://files.pythonhosted.org/packages/71/e3/19762fdfc62877ae91
02edf6342d71b28fbfd9dea3d2f96a882ce099b03f/Keras_Applications-1.0.8-py3-none
-any.whl) (50kB)
     |████████████████████████████████| 51kB 7.7MB/s
Requirement already satisfied: pyyaml in /usr/local/lib/python3.7/dist-packa
ges (from keras==2.3.1) (3.13)
Requirement already satisfied: scipy>=0.14 in /usr/local/lib/python3.7/dist-
packages (from keras==2.3.1) (1.4.1)
Installing collected packages: keras-applications, keras
  Found existing installation: Keras 2.4.3
    Uninstalling Keras-2.4.3:
      Successfully uninstalled Keras-2.4.3
Successfully installed keras-2.3.1 keras-applications-1.0.8
Collecting tensorflow_io
  Downloading https://files.pythonhosted.org/packages/88/73/a7e5eaf7d55bcf46
fe99800c39b21590351a7f4c348eac34762d4d023c1c/tensorflow_io-0.17.1-cp37-cp37m
-manylinux2010_x86_64.whl (https://files.pythonhosted.org/packages/88/73/a7e
5eaf7d55bcf46fe99800c39b21590351a7f4c348eac34762d4d023c1c/tensorflow_io-0.1
7.1-cp37-cp37m-manylinux2010_x86_64.whl) (25.4MB)
     |████████████████████████████████| 25.4MB 1.6MB/s
Requirement already satisfied: tensorflow<2.5.0,>=2.4.0 in /usr/local/lib/py
thon3.7/dist-packages (from tensorflow_io) (2.4.1)
Requirement already satisfied: h5py~=2.10.0 in /usr/local/lib/python3.7/dist
-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (2.10.0)
Requirement already satisfied: tensorflow-estimator<2.5.0,>=2.4.0 in /usr/lo
cal/lib/python3.7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_i
o) (2.4.0)
Requirement already satisfied: termcolor~=1.1.0 in /usr/local/lib/python3.7/
dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.1.0)
Requirement already satisfied: absl-py~=0.10 in /usr/local/lib/python3.7/dis
t-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (0.12.0)
Requirement already satisfied: opt-einsum~=3.3.0 in /usr/local/lib/python3.
7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (3.3.0)
Requirement already satisfied: typing-extensions~=3.7.4 in /usr/local/lib/py
thon3.7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (3.7.4.
3)
Requirement already satisfied: protobuf>=3.9.2 in /usr/local/lib/python3.7/d
ist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (3.12.4)
Requirement already satisfied: wheel~=0.35 in /usr/local/lib/python3.7/dist-
```

```
packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (0.36.2)
Requirement already satisfied: wrapt~=1.12.1 in /usr/local/lib/python3.7/dis
t-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.12.1)
Requirement already satisfied: google-pasta~=0.2 in /usr/local/lib/python3.
7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (0.2.0)
Requirement already satisfied: gast==0.3.3 in /usr/local/lib/python3.7/dist-
packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (0.3.3)
Requirement already satisfied: tensorboard~=2.4 in /usr/local/lib/python3.7/
dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (2.4.1)
Requirement already satisfied: six~=1.15.0 in /usr/local/lib/python3.7/dist-
packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.15.0)
Requirement already satisfied: grpcio~=1.32.0 in /usr/local/lib/python3.7/di
st-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.32.0)
Requirement already satisfied: numpy~=1.19.2 in /usr/local/lib/python3.7/dis
t-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.19.5)
Requirement already satisfied: keras-preprocessing~=1.1.2 in /usr/local/lib/
python3.7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.1.
2)
Requirement already satisfied: flatbuffers~=1.12.0 in /usr/local/lib/python
3.7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.12)
Requirement already satisfied: astunparse~=1.6.3 in /usr/local/lib/python3.
7/dist-packages (from tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.6.3)
Requirement already satisfied: setuptools in /usr/local/lib/python3.7/dist-p
ackages (from protobuf>=3.9.2->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (56.
0.0)
Requirement already satisfied: tensorboard-plugin-wit>=1.6.0 in /usr/local/l
ib/python3.7/dist-packages (from tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0-
>tensorflow_io) (1.8.0)
Requirement already satisfied: markdown>=2.6.8 in /usr/local/lib/python3.7/d
ist-packages (from tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_i
o) (3.3.4)
Requirement already satisfied: google-auth-oauthlib<0.5,>=0.4.1 in /usr/loca
l/lib/python3.7/dist-packages (from tensorboard~=2.4->tensorflow<2.5.0,>=2.
4.0->tensorflow_io) (0.4.4)
Requirement already satisfied: requests<3,>=2.21.0 in /usr/local/lib/python
3.7/dist-packages (from tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorfl
ow_io) (2.23.0)
Requirement already satisfied: werkzeug>=0.11.15 in /usr/local/lib/python3.
7/dist-packages (from tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow
_io) (1.0.1)
Requirement already satisfied: google-auth<2,>=1.6.3 in /usr/local/lib/pytho
n3.7/dist-packages (from tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorf
low_io) (1.28.1)
Requirement already satisfied: importlib-metadata; python_version < "3.8" in
/usr/local/lib/python3.7/dist-packages (from markdown>=2.6.8->tensorboard~=
2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (3.10.1)
Requirement already satisfied: requests-oauthlib>=0.7.0 in /usr/local/lib/py
thon3.7/dist-packages (from google-auth-oauthlib<0.5,>=0.4.1->tensorboard~=
2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.3.0)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /u
sr/local/lib/python3.7/dist-packages (from requests<3,>=2.21.0->tensorboard~
=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (1.24.3)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.
7/dist-packages (from requests<3,>=2.21.0->tensorboard~=2.4->tensorflow<2.5.
0,>=2.4.0->tensorflow_io) (2020.12.5)
Requirement already satisfied: idna<3,>=2.5 in /usr/local/lib/python3.7/dist
-packages (from requests<3,>=2.21.0->tensorboard~=2.4->tensorflow<2.5.0,>=2.
4.0->tensorflow_io) (2.10)
Requirement already satisfied: chardet<4,>=3.0.2 in /usr/local/lib/python3.
7/dist-packages (from requests<3,>=2.21.0->tensorboard~=2.4->tensorflow<2.5.
0,>=2.4.0->tensorflow_io) (3.0.4)
```

```
Requirement already satisfied: cachetools<5.0,>=2.0.0 in /usr/local/lib/pyth
on3.7/dist-packages (from google-auth<2,>=1.6.3->tensorboard~=2.4->tensorflo
w<2.5.0,>=2.4.0->tensorflow_io) (4.2.1)
Requirement already satisfied: rsa<5,>=3.1.4; python_version >= "3.6" in /us
r/local/lib/python3.7/dist-packages (from google-auth<2,>=1.6.3->tensorboard
~=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (4.7.2)
Requirement already satisfied: pyasn1-modules>=0.2.1 in /usr/local/lib/pytho
n3.7/dist-packages (from google-auth<2,>=1.6.3->tensorboard~=2.4->tensorflow
<2.5.0,>=2.4.0->tensorflow_io) (0.2.8)
Requirement already satisfied: zipp>=0.5 in /usr/local/lib/python3.7/dist-pa
ckages (from importlib-metadata; python_version < "3.8"->markdown>=2.6.8->te
nsorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (3.4.1)
Requirement already satisfied: oauthlib>=3.0.0 in /usr/local/lib/python3.7/d
ist-packages (from requests-oauthlib>=0.7.0->google-auth-oauthlib<0.5,>=0.4.
1->tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (3.1.0)
Requirement already satisfied: pyasn1>=0.1.3 in /usr/local/lib/python3.7/dis
t-packages (from rsa<5,>=3.1.4; python_version >= "3.6"->google-auth<2,>=1.
6.3->tensorboard~=2.4->tensorflow<2.5.0,>=2.4.0->tensorflow_io) (0.4.8)
Installing collected packages: tensorflow-io
Successfully installed tensorflow-io-0.17.1
```

In [3]:

```
import time
import os
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from tqdm import tqdm_notebook as tqdm
!pip install pydicom
import pydicom
from pydicom import dcmread
from sklearn.utils import resample # Handle Imbalance
import pathlib
import PIL
import tensorflow_io as tfio
import tensorflow as tf
import keras
keras.backend.set_image_data_format('channels_last')
from keras import backend as K
from tensorflow.keras.models import Model
from tensorflow.keras.losses import binary_crossentropy
```

```
Collecting pydicom
  Downloading https://files.pythonhosted.org/packages/f4/15/df16546bc59bfca3
90cf072d473fb2c8acd4231636f64356593a63137e55/pydicom-2.1.2-py3-none-any.whl
 (https://files.pythonhosted.org/packages/f4/15/df16546bc59bfca390cf072d473f
b2c8acd4231636f64356593a63137e55/pydicom-2.1.2-py3-none-any.whl) (1.9MB)
     |████████████████████████████████| 1.9MB 4.4MB/s
Installing collected packages: pydicom
Successfully installed pydicom-2.1.2

Using TensorFlow backend.
```

In [4]:

```
os.chdir('/gdrive/MyDrive/Image_Segementation_CS2/')
```

In [5]:

```
# read csv file
df_main  = pd.read_csv('Main_CS2_SIIM_All.csv')
df_downsampled  = pd.read_csv('Main_CS2_SIIM.csv')
```

In [6]:

```
df_main.head()
```

Out[6]:

	UID	Encoded_pixel	Path
0	1.2.276.0.7230010.3.1.4.8323329.1000.151787516...	-1	siim/dicom-images-train/1.2.276.0.7230010.3.1....
1	1.2.276.0.7230010.3.1.4.8323329.10000.15178752...	-1	siim/dicom-images-train/1.2.276.0.7230010.3.1....
2	1.2.276.0.7230010.3.1.4.8323329.10001.15178752...	-1	siim/dicom-images-train/1.2.276.0.7230010.3.1....
3	1.2.276.0.7230010.3.1.4.8323329.10002.15178752...	-1	siim/dicom-images-train/1.2.276.0.7230010.3.1....
4	1.2.276.0.7230010.3.1.4.8323329.10003.15178752...	-1	siim/dicom-images-train/1.2.276.0.7230010.3.1....

MODEL

UNET - ChexNet as Bonebone

In [7]:

```
# Metrics
def dice_coeff(actual,predicted,smooth=1):
  Actual = K.flatten(actual)
  Predict = K.flatten(predicted)
  intersection  = K.sum(Actual *Predict)
  return ((2.* intersection + smooth) / (K.sum(Actual) +K.sum(Predict) +smooth))
```

In [8]:

```
Segmentation_model   = tf.keras.models.load_model('new_model_save_test/best_models_Unet_Che
classification_model = tf.keras.models.load_model('best_models_classification.h5') # Loadin
```

In [9]:

```
# Original Model size (Before Qunatizsation)
file_size = os.stat('new_model_save_test/best_models_Unet_ChexNet.hdf5')
print('File Size in MD - {}'.format(file_size.st_size /(1024*1024)))
```

```
File Size in MD - 125.38428497314453
```

In [14]:

```
converter = tf.lite.TFLiteConverter.from_keras_model(Segmentation_model)
tflite_model = converter.convert()
```

```
INFO:tensorflow:Assets written to: /tmp/tmpel1ox_ur/assets
```

In [15]:

```
tflite_models_dir = pathlib.Path("/tmp/siim_tflite_models/")
tflite_models_dir.mkdir(exist_ok=True, parents=True)
```

In [16]:

```
tflite_model_file = tflite_models_dir/"siim_model.tflite"
tflite_model_file.write_bytes(tflite_model)
```

Out[16]:

```
48311532
```

In [17]:

```
file_size = os.stat(tflite_models_dir/"siim_model.tflite")
print('File Size in MD - {}'.format(file_size.st_size /(1024*1024)))
```

```
File Size in MD - 46.07346725463867
```

After Quantizsation , Model reduced from 125 MB to 46MB

In [18]:

```
converter.optimizations = [tf.lite.Optimize.DEFAULT]
converter.target_spec.supported_types = [tf.float16]
```

In [19]:

```
tflite_fp16_model = converter.convert()
tflite_model_fp16_file = tflite_models_dir/"siim_model_quant_f16.tflite"
tflite_model_fp16_file.write_bytes(tflite_fp16_model)
```

```
INFO:tensorflow:Assets written to: /tmp/tmpq7sl9ijw/assets

INFO:tensorflow:Assets written to: /tmp/tmpq7sl9ijw/assets
```

Out[19]:

```
24255840
```

In [20]:

```
interpreter = tf.lite.Interpreter(model_path=str(tflite_model_file))
interpreter.allocate_tensors()
```

In [21]:

```
interpreter_fp16 = tf.lite.Interpreter(model_path=str(tflite_model_fp16_file))
interpreter_fp16.allocate_tensors()
```

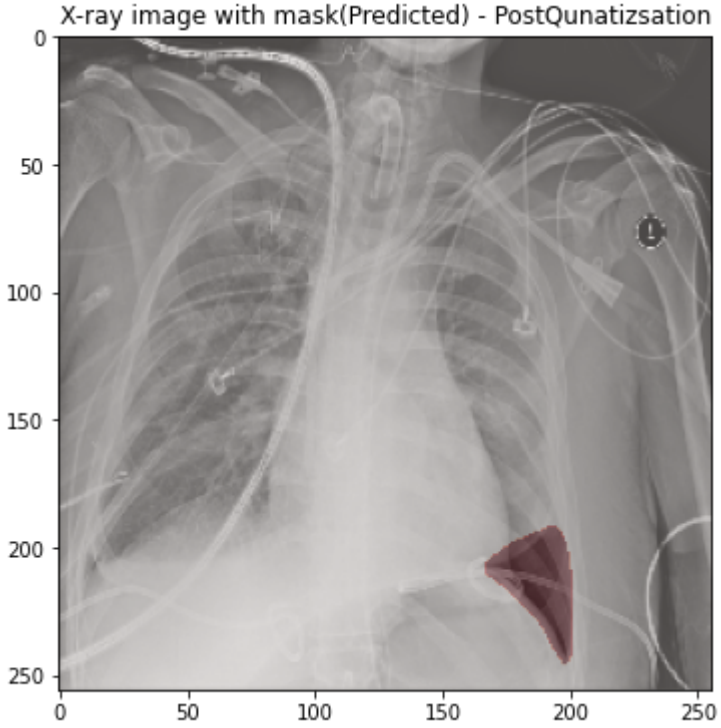
In [34]:

```
#Function-  Classification_Segmentation   --->
'''Here we are doing two actions, First we predicting whether given image has affected by p
    If Yes, Display X-ray with highlighted affected part.
    If No, Display image as it is. 😀
'''

def Classification_Segmentation_PostQuantizsation(X):
  img = tf.io.read_file(X)
  image = tfio.image.decode_dicom_image(img, dtype=tf.uint8,color_dim=True,scale='preserve'
  image = tf.image.convert_image_dtype(image, tf.float32)#converting the image to tf.float3
  image=tf.squeeze(image,[0]) #squeezing the image because the file is of the shape(1,1024,
  b = tf.constant([1,1,3], tf.int32)
  image=tf.tile(image,b)#the image is of the shape (1024,1024,1) to make it (1024,1024,3) I
  image_1=tf.image.resize(image,size=[256,256])
  image=tf.expand_dims(image_1,axis=0)

  if classification_model.predict(image)>=0.5:
    print("Pneumothorax has been detected")
    test_image = np.expand_dims(image_1, axis=0).astype(np.float32)

    input_index = interpreter.get_input_details()[0]["index"]
    output_index = interpreter.get_output_details()[0]["index"]

    interpreter.set_tensor(input_index, test_image)
    interpreter.invoke()
    predictions = interpreter.get_tensor(output_index)
    mask=predictions[0]
    mask=(mask>0.5).astype(np.uint8)

    plt.figure(figsize=(10,6))
    plt.title("X-ray image with mask(Predicted) - PostQunatizsation")
    plt.imshow(np.squeeze(image),cmap='gray')
    plt.imshow(np.squeeze(mask),cmap='Reds',alpha=0.3)
    return plt.show()

  else:
    plt.figure(figsize=(10,6))
    print('Person is Healthy, No Pneumothorax is detected')
    plt.imshow(np.squeeze(image),cmap='gray')
    return plt.show()
```

In [38]:

```
start_time = time.time()
Classification_Segmentation_PostQuantizsation(df_downsampled['Path'][45])
print("--- %s seconds --- for execution" % (time.time() - start_time))
```

```
Pneumothorax has been detected
```




```
--- 0.5093870162963867 seconds --- for execution
```

In [ ]:

```

```

Without Quantizsation

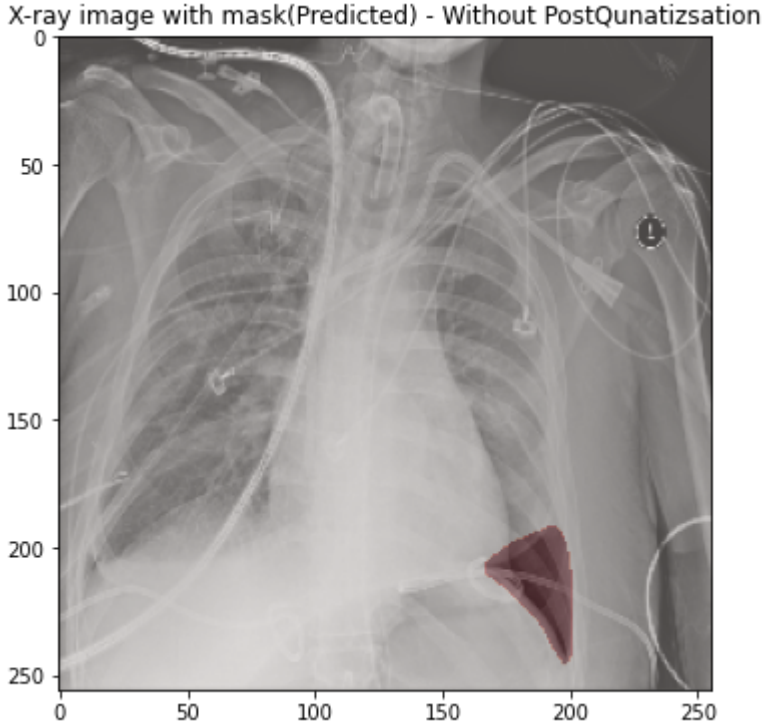
In [41]:

```
#Function-2 ---> If Function 1 predict, image having Pneumothorax, then highlight the affec
def Classification_Segmentation(X):
  img = tf.io.read_file(X)
  image = tfio.image.decode_dicom_image(img, dtype=tf.uint8,color_dim=True,scale='preserve'
  image = tf.image.convert_image_dtype(image, tf.float32)#converting the image to tf.float3
  image=tf.squeeze(image,[0]) #squeezing the image because the file is of the shape(1,1024,
  b = tf.constant([1,1,3], tf.int32)
  image=tf.tile(image,b)#the image is of the shape (1024,1024,1) to make it (1024,1024,3) I
  image=tf.image.resize(image,size=[256,256])
  image=tf.expand_dims(image,axis=0)

  if classification_model.predict(image)>=0.5:
    print("Pneumothorax has been detected")
    mask=Segmentation_model.predict(image)
    mask=(mask>0.5).astype(np.uint8)

    plt.figure(figsize=(10,6))
    plt.title("X-ray image with mask(Predicted) - Without PostQunatizsation")
    plt.imshow(np.squeeze(image),cmap='gray')
    plt.imshow(np.squeeze(mask),cmap='Reds',alpha=0.3)
    return plt.show()

  else:
    plt.figure(figsize=(10,6))
    print('Person is Healthy, No Pneumothorax is detected')
    plt.imshow(np.squeeze(image),cmap='gray')
    return plt.show()
```

In [42]:

```
start_time = time.time()
Classification_Segmentation(df_downsampled['Path'][45])
print("--- %s seconds --- for execution" % (time.time() - start_time))
```

```
Pneumothorax has been detected
```

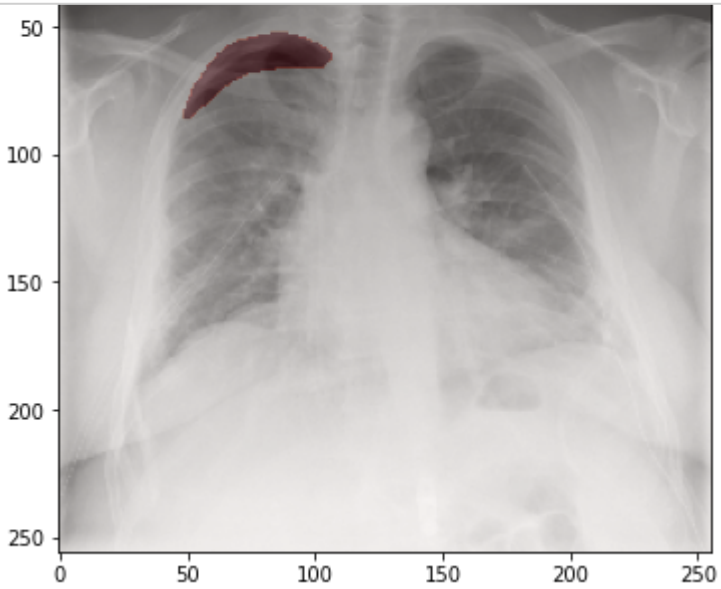



```
--- 2.4546945095062256 seconds --- for execution
```

PostQunatizsation and original model predictions - Samples

In [43]:

```
import random
random_val = random.sample(range(100), 20)

for i in random_val:
  #Call Post quantizsation
  Classification_Segmentation_PostQuantizsation(df_downsampled['Path'][i])

  #Call Without Post quantizsation
  Classification_Segmentation(df_downsampled['Path'][i])
  print('**'*80)
```




```
********************************************************************************
********************************************************************************
```

SUMMARY

1. Model size reduced from 125 MB to 45MB and model performances not affected
2. After postquantizsation, we reduced the execution time from 2.4 seconds to 0.5 seconds and result also similar