Sensitive Data Identification

Sanoop Mallissery

Sensitive Data Identification

- ➢ Medical Images from Scan results contains Data with sensitive information such as

{Name, Age, Gender, Patient ID, Date of Birth, SSN etc}

- ➢ The Proposed system uses Mask RCNN for Sensitive data Identification

Mask Regional-Convolutional Neural Network (Mask R-CNN)

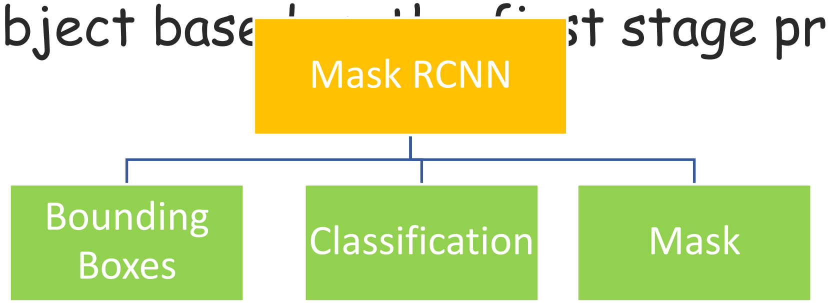
- ➢ A deep neural network aimed to solve instance segmentation problem.
- ➢ Here it is used to perform sensitive data detection on medical images.

First stage:

It uses Region Proposal Network (RPN) to generate those proposals that contain higher probability of sensitive data

Second stage:

It Predicts the class of the object, refines the bounding box and generates a mask in pixel level of the object based on the first stage proposals



Mask RCNN

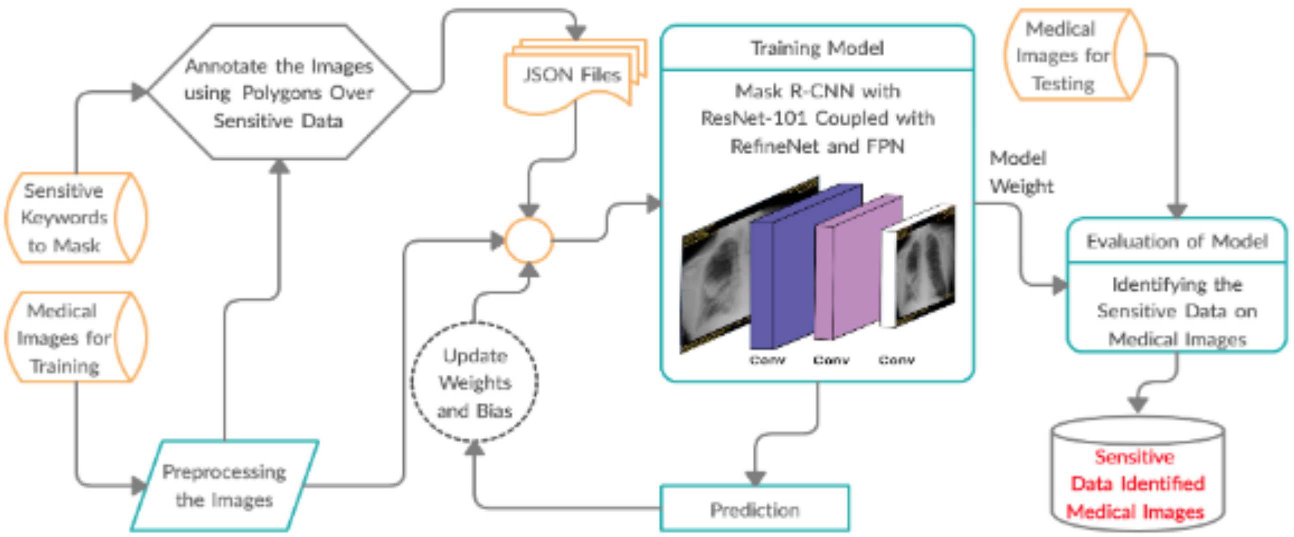
Bounding Boxes

Classification

Mask

System work Model & Phases

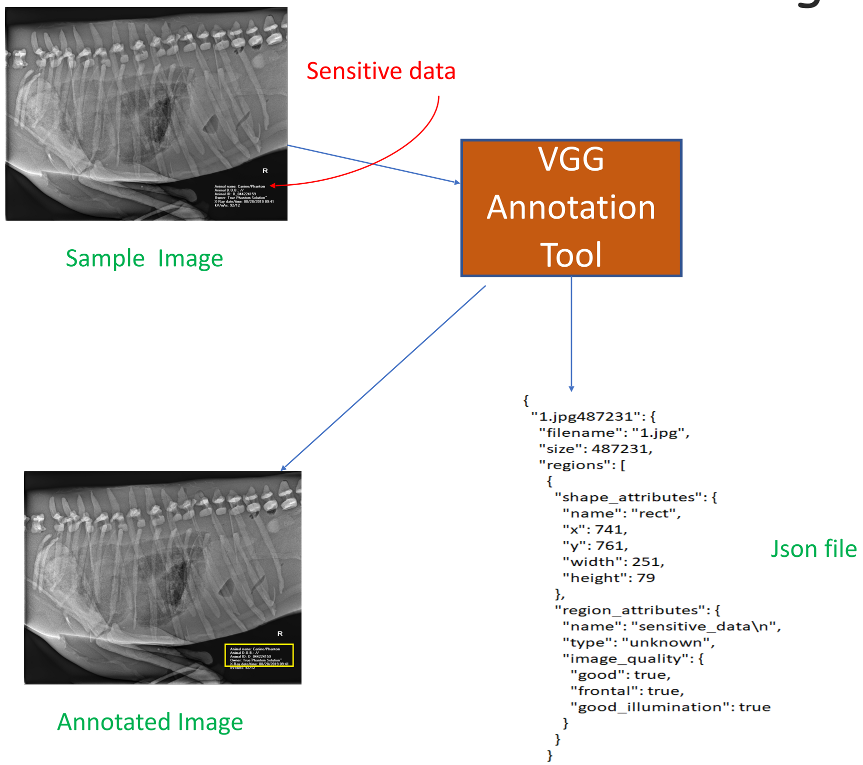
System Model for various phases of Automated Identification of Sensitive data


Annotate the Images using Polygons Over Sensitive Data
JSON Files
Training Model
Mask R-CNN with ResNet-101 Coupled with RefineNet and FPN
Conv
Conv
Conv
Medical Images for Testing
Model Weight
Evaluation of Model
Identifying the Sensitive Data on Medical Images
Sensitive Keywords to Mask
Medical Images for Training
Update Weights and Bias
Preprocessing the Images
Prediction
Sensitive Data Identified Medical Images

Step 1: Data Collection & Pre-processing

- ➢80% of medical images are used to train the model, and 20% of medical images are used for evaluating the model.
- ➢All images are resized to {1024 ×1024} pixels.

- ➢ VGG Image Annotation tool is used to annotate the sensitive data of the images using a polygon/ bounding box
- ➢ One Json file is generated for one folder and this is been fed to the CNN model for Training.



```
{
 "1.jpg487231": {
  "filename": "1.jpg",
  "size": 487231,
  "regions": [
   {
    "shape_attributes": {
     "name": "rect",
     "x": 741,
     "y": 761,
     "width": 251,
     "height": 79
    },
    "region_attributes": {
     "name": "sensitive_data\n",
     "type": "unknown",
     "image_quality": {
      "good": true,
      "frontal": true,
      "good_illumination": true
     }
    }
   }
```

System work Model & Phases

Mask RCNN with Resnet 101 & RefineNet + FPN is chosen as the model for training.

- The "classify" feature of ResNet-101 is used to classify the images which represent sensitive data in one class
- The data bounded by box/ polygon is the ROI (Region of Interest)

- Passing through the backbone network, the image is converted from 1024x1024px x 3 (RGB) to a feature map of shape 32x32x2048

- The bottom-up and top-down pathway connections of FPN allow the features to access the features of lower and higher-level at each layer

- Implementation of training and testing is carried out in python using the libraries such as Keras and Tensor Flow

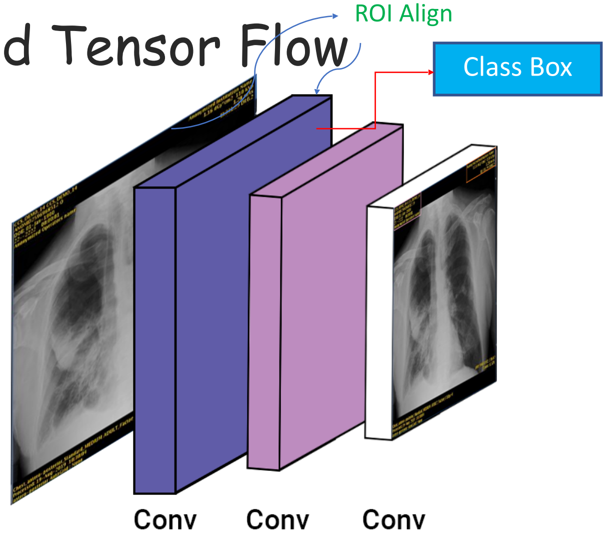


Illustration of Mask RCNN

Evaluation Results

The training of model is started with the initialization of some random values to Weight (W)and bias (b) to predict the output with those values.
Initial weight of the model is set to mask_rcnn_coco.h5

Batch size : 1 image/GPU
Epoch : 300
Steps : 100

The values of W and b are adjusted on comparison with the output predicted model so as to achieve more accuracy

The 20% of total Images is evaluated against the model and the result is as follows