TABLE DETECTION FROM SCANNED DOCUMENT IMAGES

EE 691 - R&D Project

Ayush Munjal, 19D070014
Department of Electrical Engineering
Indian Institute of Technology Bombay, Mumbai

Guided By Prof. Rajbabu Velmurugan & Prof. Biplab Banerjee

Abstract

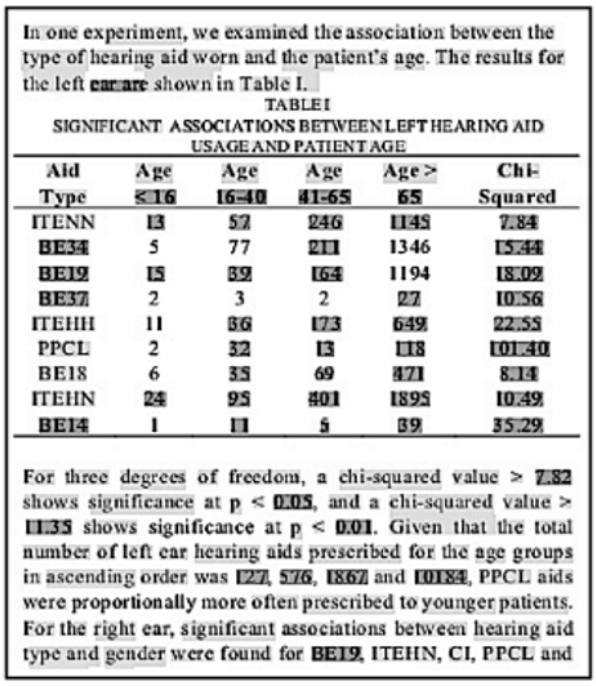
This document discusses the task of extracting tabular data from scanned document images. This task can be divided into two sub-tasks, recognising the table positions in the image, extracting thr tabular data. We explore Encoder-Decoder Transformer architecture for this task.

Contents

1 Introduction

With the increasing amounts of data, extracting reliable data from documents is becoming an important task. Images are one of the largest sources of data. Tables are one of the essential sources of data in images. Extracting tables from images is still a big challenge because we have to extract the table boundary in the table's image, rows, and column information and extract the data from a table.

We use TableNet model architecture, a deep-learning model for table detection and structure recognition. We will test this model with different pre-trained models. Further, we also apply a post-processing technique to improve the accuracy. One such example of table extraction from image is as shown below.



In one experiment, we examined the association between the type of hearing aid worn and the patient's age. The results for the left earare shown in Table I.

TABLE I
SIGNIFICANT ASSOCIATIONS BETWEEN LEFT HEARING AID USAGE AND PATIENT AGE

Aid Type	Age ≤ 16	Age 16–40	Age 41–65	Age > 65	Chi-Squared
ITENN	13	57	246	1145	7.84
BE34	5	77	211	1346	15.44
BE19	15	89	164	1194	18.09
BE37	2	3	2	27	10.56
ITEHH	11	86	173	649	22.55
PPCL	2	32	13	118	101.40
BE18	6	35	69	471	8.14
ITEHN	24	95	401	1895	10.49
BE14	1	11	5	39	35.29

For three degrees of freedom, a chi-squared value > 7.82 shows significance at p < 0.05, and a chi-squared value > 11.35 shows significance at p < 0.01. Given that the total number of left ear hearing aids prescribed for the age groups in ascending order was 127, 576, 1867 and 10184, PPCL aids were proportionally more often prescribed to younger patients. For the right ear, significant associations between hearing aid type and gender were found for BE19, ITEHN, CI, PPCL and

Figure 1: Tabular data present in an image

2 TableNet

Table extraction task can be divided into two parts:

- Table detection: Detecting co-ordinates of the table boundary
- Structure recognition: Segmentation of individual rows and columns

TableNet exploits the interdependence of these two problems. It uses an Encoder-Decoder model. It consists of a single common encoder using a base pre-trained model and two decoders one for predicting the column area (column mask) and one for table area (table mask). In this paper, VGG19 was used as the pre-trained model. The architecture of the model is as shown below.

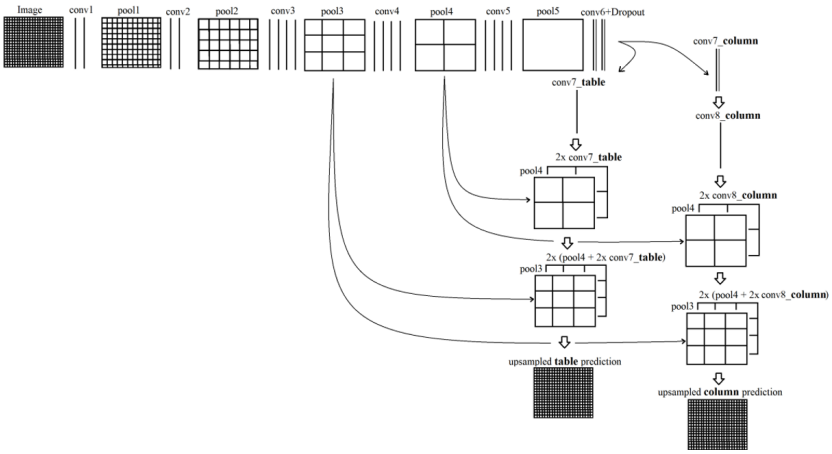
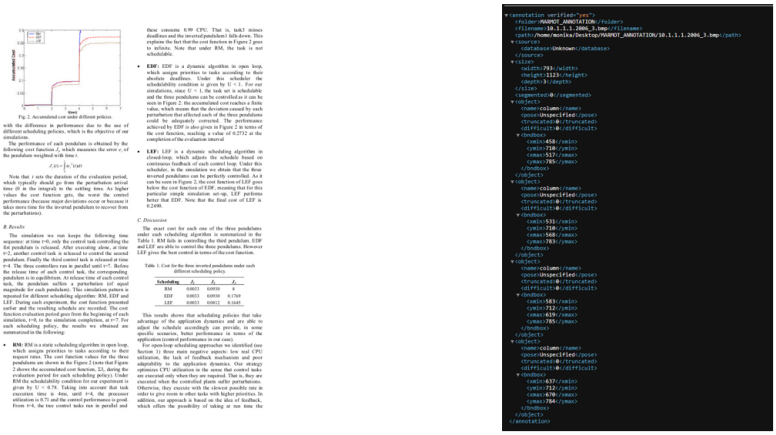


Figure 2: TableNet model architecture

3 Dataset

The Marmot dataset is used. It consists of 1016 documents in Chinese and English. These documents consist of conference and journal papers with great variety in page layout and table styles. It has 509 English annotated documents. We use English documents only. The image data is stored in a *.bmo* file, and the bounding box coordinates are stored in *.xml* file as shown below.



Document Image XML file

Figure 3: Image and XML file for that image containing table and column mask coordinates

Figure 4: Corresponding table and column mask

4 Model architecture

We use the TableNet architecture. It consists of a base pre-trained model used as encoder, separate decoder layers for table and column mask. Both decoder layers consists of two alternate convolution and pooling layers. Any image classifier model can be used as encoder. We use three pre-trained models:

- ResNet51
- Vgg19
- DenseNet121

5 Experiments and Results

We divide the dataset, with 444 samples for training and 50 samples for testing. We train the model for 500 epochs. The sparse Categorical Cross Entropy loss function is used for both table and column mask loss, with the same weights for both losses. Adam optimizer is used.

5.1 Comparison of three models

Three models were trained using different pre-trained models, as discussed earlier. The figure below shows the training loss with the number of epochs for these three models for same learning rate.

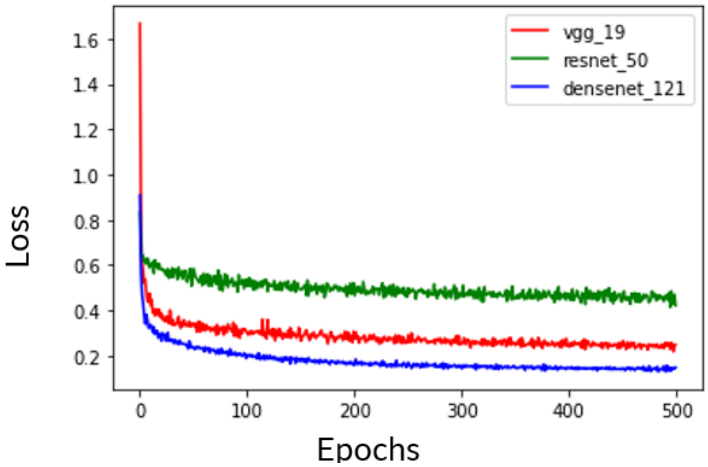


Figure 5: Training loss for different pre-trained models with number of epochs

5.2 Hyperparameter tuning for DenseNet121

As we observed from the previous result that DenseNet121 performs far better than other models, we perform hyperparameter tuning for this model by changing the learning rate and weight decay. Results for both these hyperparameters are shown below.

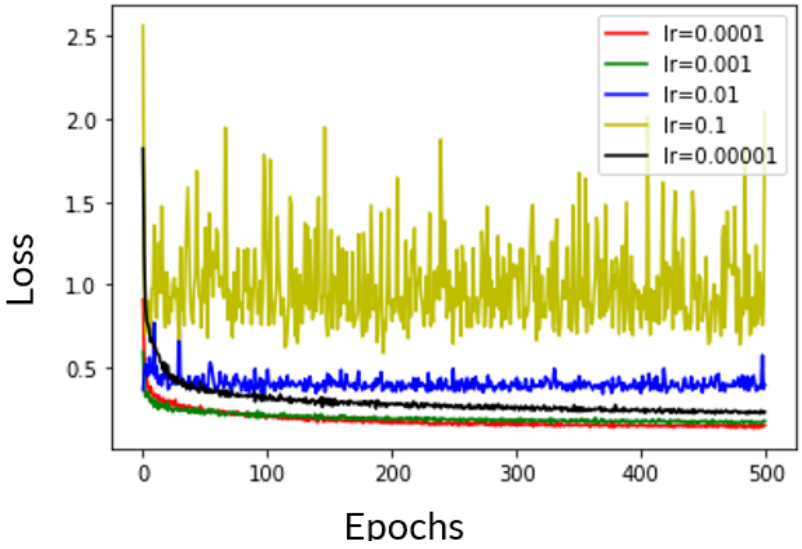


Figure 6: Training loss for different learning rates for DenseNet121 as pre-trained model

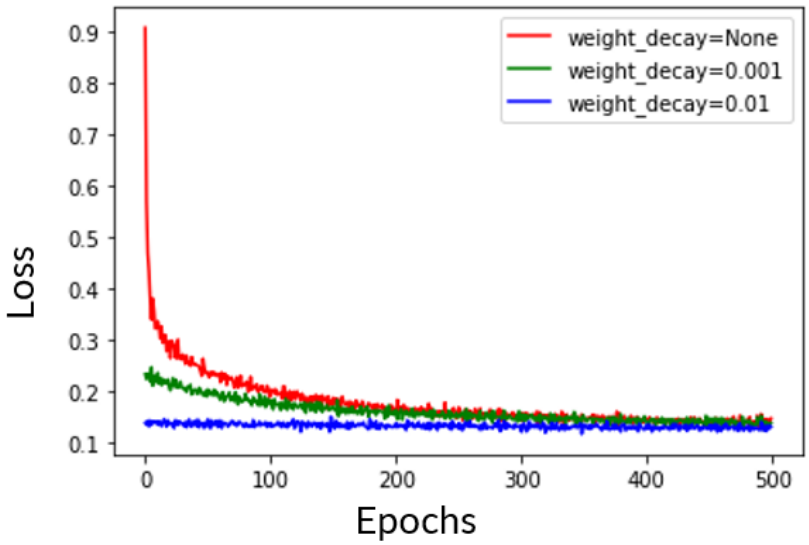


Figure 7: Training loss for different weight decay values for DenseNet121 as pre-trained model

5.3 IoU scores & F1 values

IoU is the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the ground truth. IoU score is a better metric for image segmentation. IoU scores for three pre-trained models on test dataset for table and column mask are shown in the table below.

Model	IoU score for table mask	IoU score for column mask
DenseNet121	0.84	0.72
VGG19	0.67	0.50
ResNet50	0.36	0.18

Figure 8: Average IoU scores for test dataset for VGG19, DenseNet121, and ResNet50

We further evaluate these three models on some test images and compare their IoU values below.

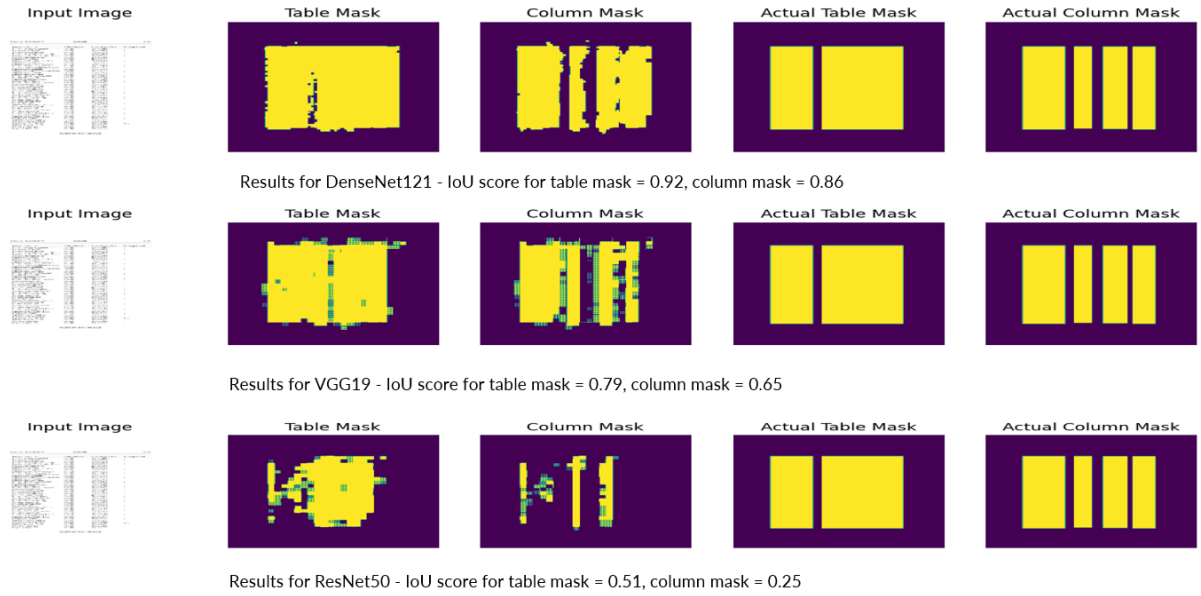


Figure 9: Comparison of the results obtained for these three models with IoU values

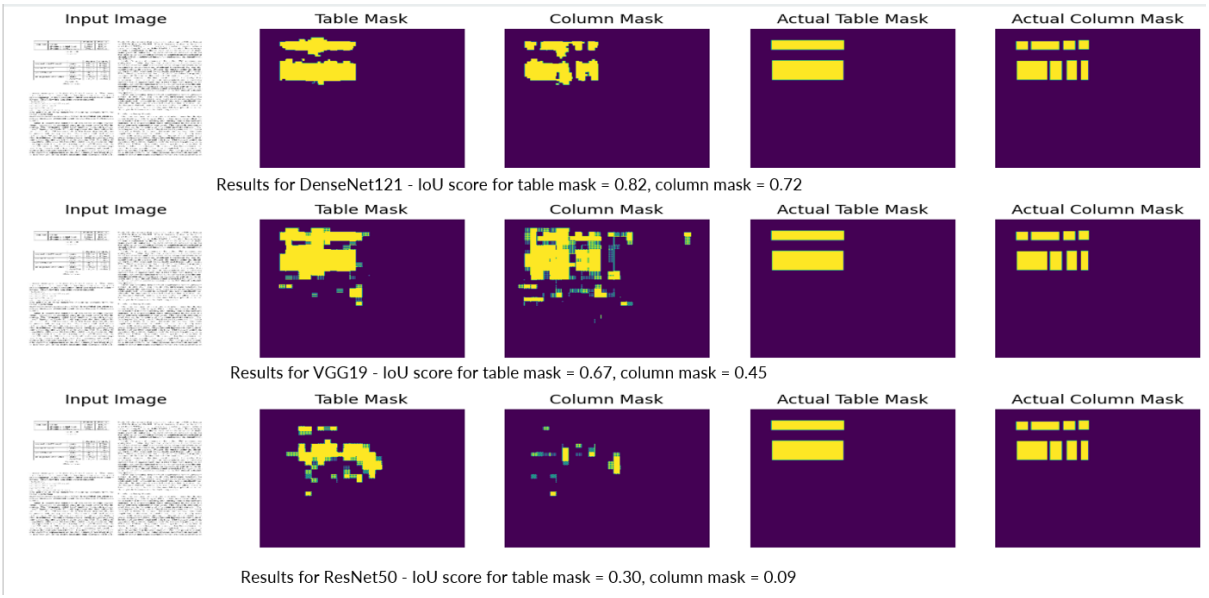


Figure 10: Comparison of the results obtained for these three models with IoU values

We also calculated average F1 values for these three models for the testing dataset. The values obtained are as shown in the table below

Model	Table mask F1	Column mask F1
DenseNet121	0.91	0.84
VGG19	0.77	0.62
ResNet50	0.54	0.37

Figure 11: Average F1 scores for test dataset for VGG19, DenseNet121, and ResNet50

5.4 Results on rotated images

To check the robustness of our model, we also rotated some sample images and checked the results produced with the three models on these images, as shown below.

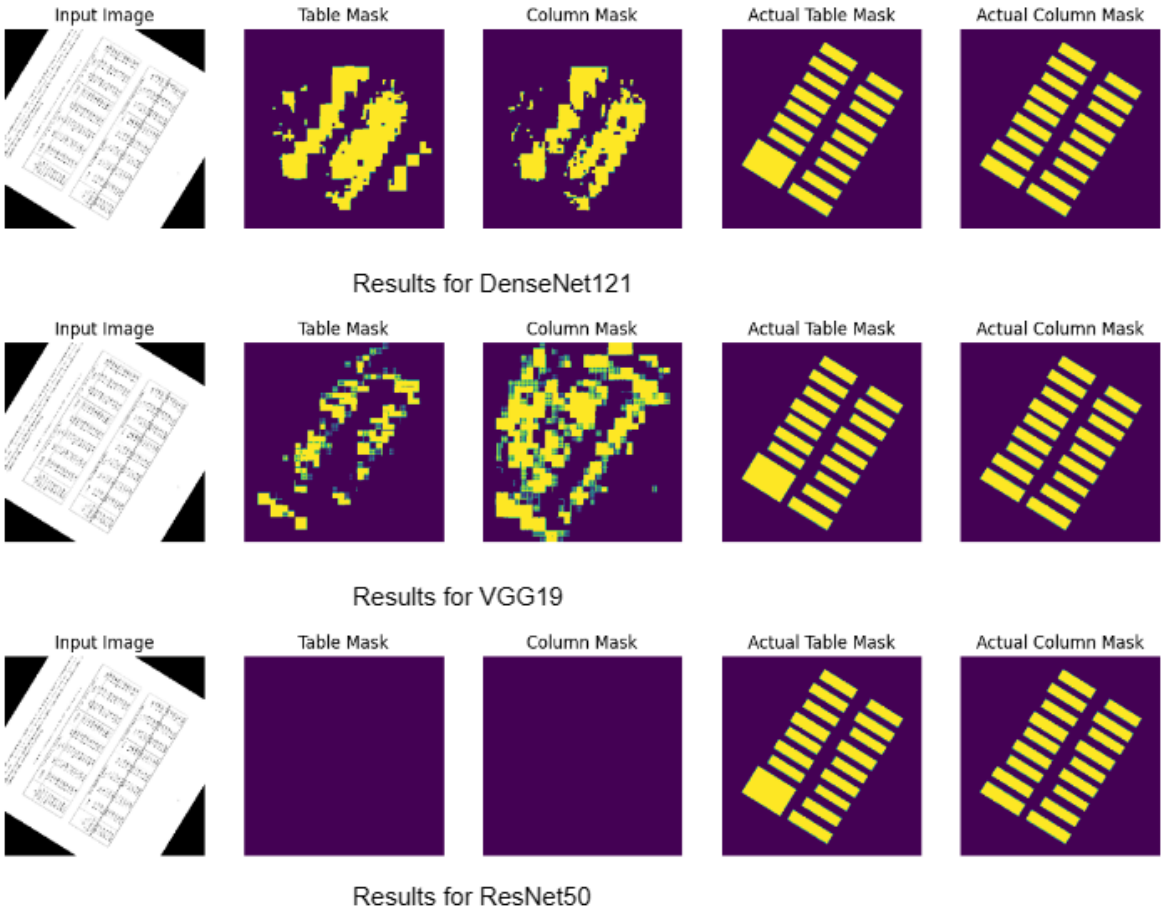


Figure 12: Comparison of the results obtained for these three models on rotated images

6 Conditional Random Fields (CRF) post processing technique

Now, we employ a post-processing technique to further improve the accuracy of our model. One popular post-processing technique for image segmentation tasks is Conditional Random Fields post-processing. It smoothens the image segmentation output by using the neighbor information.

6.1 CRF

CRF model exploits the dependencies within input domains. There are different types of CRF models depending on how many neighbors to consider:

- Linear
- Grid
- Dense

We will use a dense CRF model. CRFs often used for sequence pattern recognition tasks. It incorporates subjective and non-independent features of information.

6.2 Implementation

No direct implementation of CRF available for image segmentation task. We use CRF-RNN repository used which provides an implementation of CRF layer for image segmentation tasks. It provides an easy-to-use implementation of the CRF layer. We use DenseNet121 with optimal learning rate and weight decay as obtained in previous section. We only add a CRF layer for the table mask. We added a CRF layer after the table output and froze the weights of the original model. Sparse Categorical Cross entropy loss is used again. Adam optimizer is used again. The number of iterations is set to 100.

6.3 Results

The results obtained with CRF layer is shown in the table below.

Learning rate	Table mask accuracy
0.01	0.9663
0.001	0.9673
0.0001	0.9658

Figure 13: Table mask accuracy values with CRF layer

We further compare the CRF model results with the previous model (with any post processing scheme), the results obtained are as shown below.

Figure 14: Comparison of table mask produced with and without CRF with IoU values

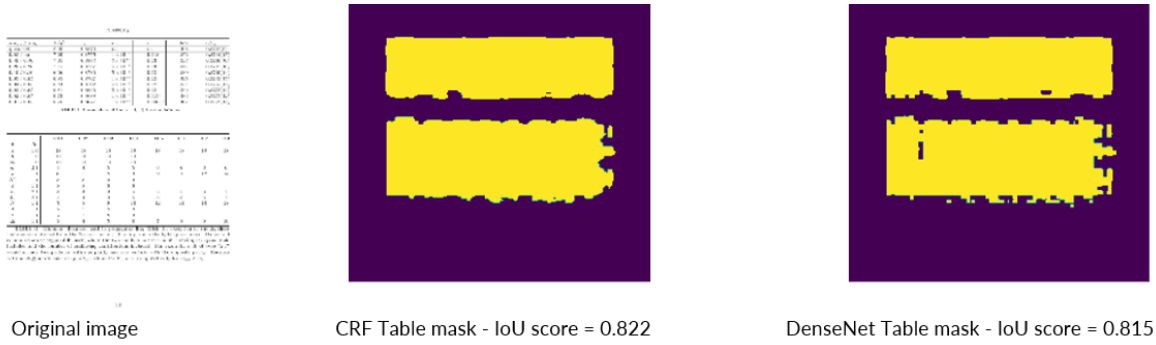


Figure 15: Comparison of table mask produced with and without CRF with IoU values

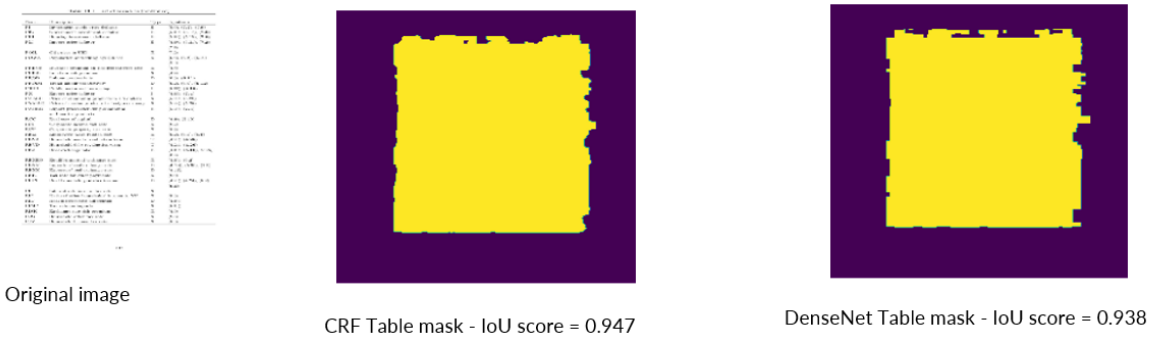


Figure 16: Comparison of table mask produced with and without CRF with IoU values

We can observe that IoU values have increased by 0.01 for each image, although this change is small the results produced by CRF model are much more smooth than original results.

7 Conclusion

DenseNet121 provides us with much better results than using VGG19. We also employed a post-processing technique to further improve the IoU values. However, CRF post-processing proved little helpful as there were only slight improvements in IoU values. Thus, we can look for more image segmentation post-processing techniques. It is also observed that table mask results are satisfactory while column mask results are not. Therefore, we will have to improve the column mask prediction. We can use different weight values for table mask and column mask prediction in the loss function to observe the improvement in column mask. We can also use IoU loss instead of Cross Entropy loss or a mixture of both.

References

[1] Shubham Paliwal, Vishwanath D, Rohit Rahul, Monika Sharma, & Lovekesh Vig. (2020). TableNet: Deep Learning model for end-to-end Table detection and Tabular data extraction from Scanned Document Images

[2] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, & Philip H. S. Torr (2015). Conditional Random Fields as Recurrent Neural Networks

[3] Dhawan, A., Bodani, P., & Garg, V. (2019). Post Processing of Image Segmentation using Conditional Random Fields

[4] crfasrnn-keras: CRF-RNN Keras/Tensorflow version