Text Segmentation for Optical Character Recognition

Stage-II Mid-Sem Report

Submitted by

Vasu Sharma 111803127

Under the guidance of

Mr. Mandar Waghmode

IE4 Labs,Axis Technical Group

and

Prof. Suraj Swawant

College Of Engineering, Pune

DEPARTMENT OF COMPUTER ENGINEERING AND INFORMATION TECHNOLOGY, COLLEGE OF ENGINEERING, PUNE

March, 2022

Abstract

Optical Character Recognition(OCR) is widely used across many products and services offered by the company. Text localization/segmentation and text recognition form most important parts of any optical character recognition pipeline. The present text segmentation techniques in place perform well for machine printed documents but are not as effective for documents with handwritten content. We are working on improving the deep learning models for text localisation/segmentation part of the pipeline. We so far have explored 2 network architectures based on the literature surveyed - ARU Net and CRAFT. ARU Net was originally designed to be used on historic documents while CRAFT was designed for localization generic "text in the wild" scenario. In the data prepartion and pre-processing step - we prepared training data consisting of a wide variety of documents. To improve the results, we introduced image augmentations and modified the networks under study as per our observations. We introduced some post processing techniques to get better end-to-end results for the pipeline. We are able to get good line-level segmentation using ARU Net combined with some post-processing. For CRAFT, with a fully supervised training setup, we have been able achieve good segmentation results for machine print documents but improvements are needed for certain subset of our data and especially so for handwritten ones. Further experiments with semi-supervise training of modified CRAFT-like are in progress to further address these shortcomings.

Keywords:

ARU Net - Attention Residual U-Net

CRAFT - Character Region Awareness for Text Detection

Contents

Introduction

Optical characeter recognition (OCR) generally refers to the process of recognizing the text in images of scanned papers, digital photographs, computer screens etc. The text can be machine printed or handwritten. OCR engines find a lot of applications like document archival, indexing, information extraction etc in the fields like medical records, financial and banking records. These usecases demand high accuracy/precision from OCR engines especially in cases of sensitive information.

There are serveral optical character recognition engines available - both community driven ie. free ones as well as subscription based. Some of the most popular Optical Character Recognition Engines include Tesseract, OpenCV, Omnipage, Google OCR and Microsoft's Aforge. Tesseract is a very popular engine available on various operating systems. It is a free software, released under the Apache License. In many cases, Tessaract's results are not as accurate as some other commercial solutions available in the market. Also, it doesn't do well with images affected by artifacts including partial occlusion, distorted perspective, and complex background. And most importantly detecting handwriting.

Several commercial tools like Omnipage, Abby, Google offer relatively good Optical Character Recogition on machine print images but relatively mixed results on handwritten text or in situations where the text is not perfectly aligned. Also in some cases, the user interface is not so intuitive for naive users. Several other subscription or paid softwares offering Optical Character Recognition are designed to be very generic and do not effectively address the needs and use cases that the company targets.

Over the years there have been extensive research proposing various deep learning based solutions for OCR. There have been solutions proposing the use of Convolutional Neural Networks (CNN) with Recurrent Neural Networks (RNN) for text recognition. While some other solutions propose transformer architecture for the task. Work has also been done applying LSTMs with attention mechanism to recognise characters better. Even with the application of these various approaches, handling of various image artifacts and handwritten characters still pose many challenging situations.

The work done here mainly pertains to the text segmentation part of a typical OCR pipeline and specifically tries to address handwritten text and certain complex cases of machine print data. Several approaches have been discussed in the subsequent sections. Segmentation here refers to detecting the regions consiting of words and lines out of the complete image that narrows down the problem as well as helps to do better OCR by reducintg errors in subsequent steps of of word-level recognition. Based on our literature survey - we narrowed down on two differnt approaches for text segmentation - CRAFT and ARU-Net.

Taking a step back, problems in the Computer Vision domain can be divided into 3 categories - Image Classification, Object Detection and Image Segmentation. Of all these 3, Image Segmentation poses a completely different set of problems. It involved labeling each pixel in the input image and the fact that the images can be 1000s of pixels in height and width makes it very expensive to have an output for each pixel. But using the advantage of Convolutional Neural Networks, the problem becomes a lot more tractable. There have been efficient networks developed over the years, one of them being U-Net for tackling the problem in a less-expensive way. ARU Net and CRAFT build on top of U-Net for text extraction using segmentation of images.

Literature Survey

Efficient, Lexicon-Free OCR using Deep Learning [2019][1]: This paper tries to address the general OCR problem for both printed and scene text using segmentation free recognition methods.The approach suggested by authors uses Convolutional Neural Networks to extract latent representations of input images,thus increasing robustness to local distortions.Then the features extracted by CNNs were combined and fed into the LSTM network with a Connectionist Temporal Classification (CTC) Loss function.The authors also propose a novel data augmentation technique that improves the robustness of neural models for text recognition and the model is trained purely on syntheticcally generated documents.

U-Net: Convolutional Networks for Biomedical Image Segmentation [2015][2]: This paper addresses the per-pixel segmentation problem. The network architecture consists of several downsampling layers which aim to extract the low level features from the image and the output after this stage is an image with very high number of channels and very small height and width dimensions. This output is then passed to upsampling layers which consist of transpose convolutional operations as opposed to usual max pooling operation.There are skip connections from downsampling layers to upsampling layers as the later layers forget the spatial information in the input image which is required to do segmentation.This upsampling is followed by fully connected layers which producce the output of same dimension as input in height and width and number of layers being equal to the number of segmentation classes.

ARU-Net: A Two-Stage Method for Text Line Detection in Historical Documents [2019][5]: This paper builds on top of U-Net Architecture. The work done in the paper focuses on identifying the text lines on the documents.The authors have proposed an approach with some additions to U-Net to draw the baselines for the text lines and also to draw the line starting and ending demarcations as it can be crucial for multi columnar documents. The appoach is to combine Residual Network[4] and attention to the layers used in U-Net. As suggested by the authors, there is loss of spatial information as we go deep inside the convolutional layers in the U-net, so the ResNets improve the representation power.Furthermore, spatial attention mechanism is developed which allows the ARU-Net to focus on image content at different positions and scales. The authors then use the baselines predicted using deep learning approach to further reach at text level segmentation using some heuristics

Character Region Awareness for Text Detection [2019][6]: This paper as opposed to other scene based text detection methods based on neural networks which use explict bounding boxes for making predictions, uses character level information and affinity betweeen different characters.The network is inspired by U-Net based upsampling architecture succeeded by VGG-Net for downsampling.The output of the model are two images - one consisting of character region scores and the other consisting of affinity scores between the characters. Since, it can be a difficult to obtain character level bounding box information for ground truth data, the authors have suggested using 2 slightly different architectures -

one of them is fully supervised learning based which trained on synthetic character level data generated by the researchers while the other is semi-supervised approach where the character level ground truth is generated using the interim model learned by training on the synthetic data. The final inference for word level bounding boxes is generated using some open-cv techniques wherein connected component analysis is done on the region and affinity scores together which brings out the bounding boxes for the words.The network produces really good results but at the same time is very computationally expensive to train, especially in the semi supervises phase where pseudo ground truths are generated during the training.

READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents [2017][3]: A baseline is defined in the typographical sense as the virtual line where most characters rest upon and descenders extend below. About 2000 document images from each of 9 different European archives were collected. These documents were written between 1470 and 1930. The authors sampled 250 images from each archival collection using freely available python scripts.This results in a set of 2500 document images. After removing images due to quality as well as content issues the number reduced from 2250 to 2118. For these images the text regions as well as baselines were annotated by DigiTexx. The well-known PAGE XML3 scheme is used for storing text region and baseline information

Problem Statement and Research Objectives

To improve Optical Character Recognition for handwritten and machine print documents by doing Text detection, localisation and segmentation

Although the established techniques work well for segmentation of straight-forward machine print, images distorted due to camera angles, brightness, focus etc, containing handwritten content, containing mixed text orientation and fonts etc pose a significant challenge for text localization as well as text recognition. Errors in the earlier segmentation stage compund the problems in subsequent stages of the pipeline resulting in poor outcomes.

This problem can be addressed by using deep learning based models for text localisation, where in we identify specific areas where the text is located in the page, extract those regions and pass them to the line or word-level recognition models. The deep learning models are so developed that they are agnostic to camera angles, noise and many other distortions in the image.

As an additional improvement, we also want to get character level boundaries in the image along with word or line regions/boundaries. This is a crucial feature in reducing the human effort needed in cleaning up OCR or information extraction results, as the manual effort can now be directed to verify particular characters in the image with lower confidence.

Objectives

1. To develop deep learning based models for text detection, localisation and segmentation in the docuements of handwritten and machine print data, eventually to get better at recognizing the words and lines in the documents using already developed models.
2. To improve word and line OCR models for text recognition post text detection step, speciifically for handwritten data.
3. To get better at recognising text in the bad quality images with distortions like noise, random camera angles, brightness, different ink colour etc.
4. To be able to precisely identify character level boundaries along with recognising words in the image so as to be able to re-direct the human user to verify the characters with lower confidence.

Proposed Methodology/Solution

We went on to perform two experiments parallely for text segmentation - training models for ARU Net and CRAFT. While ARU Net produces baselines for the text, post-processing techniques are applied to reach at text-line level segmentation from the page[5]. The CRAFT architecture produces character region and affinity scores which are also passed through post-processing methods to reach at word level bounding boxes.[6]

The authors of the ARU Net trained the model mainly on handwritten historic documents. The model weight files avaialable are useful just for inference purpose while cannot be used for transfer learning. The pretrained weights work well on similar historic documents and produce descent baselines for simpler non-augmented handwritten documents but fails to find proper baselines for machine print documents with random distortions and augmentations. So, for dealing with all these problems, we came up to a custom dataset for training. The dataset we used consists of a mix of historic documents, IAM dataset and some internal confidential documents handpicked from the ones on which esablished OCR tools do very well.

The baselines for historic documents are annotated using some open source python scripts and Digi-Texx as cited in the paper mentioned in Literature Review Section[3]. The IAM database contains 13,353 images of handwritten lines of text created by 657 writers. The texts those writers transcribed are from the Lancaster-Oslo/Bergen Corpus of British English and it resembles real world handwritten data. It includes contributions from 657 writers making a total of 1,539 handwritten pages comprising of 115,320 words and is categorized as part of modern collection. The database is labeled at the sentence, line, and word levels. After running omnipage on the samples, we handpick the images for which we get accurate predictions for the lines and words and then form baselines using pre-processing methods on the word level bounding box outputs that we get from existing tools. The final dataset consisted of around 1024 images for training and 128 images for validation.

To tackle the problem of under performance on the augmented images and on the images which are not so clear, we introduced rigorous augmentations on the training data on the fly so as to match the real world data on which we have to perform inferences. We introduced elastic deformations, affine transformations, horizontal flips, random rotation, perspective transformations etc, completely replicating the real world inference data.

After collection of all the data and all the augmentations in place, we trained ARU Net from scratch without fine-tuning for around 200 epochs with 250 batch steps per epoch. The network learns to draw decent baselines even on training on a very small dataset and we get to see good baselines even after training for few epochs. The results start to saturate after 150 epochs.

The ground truths and the output of the network is a set of 3 images where in first image has baseline regions and the second image has vertical boundaries at the start and end of each baseline, particularly useful to identify the lines in multi-columnar documents.The third image is just the inverted image of the image obtained by combining first 2 images together.

After we get the baselines, we employ a post-processing technique to reach to the text segmentation. We super-impose the baselines obtained from ARU Net on top of the corresponding text in the image and as a result we connect all the words in a line with the baseline. This is followed by image cleaning and discarding some pixels based on thresholds obtained by hyper-parameter tuning and utlimately we perform connected component analysis where in we identify the regions which are connected together, which in our case are the differnet words in a line. The output from the connected component analysis is passed on to detect the minimum area rectangle which will completely surround the connected region and hence we get the bounding boxes around the words in a line.

For CRAFT based approach, we planned to train the model both using supervised learning and semi-supervised learning. As for now, the training for supervised learning method has been done. For superivised learning, we need a dataset with character level bounding boxes and word labels. We prepared custom dataset using IAM data avaialable and some handpicked proprietary images. To completely replicate the real world exmaple cases, we drew synthetic baselines below the IAM images. The final dataset consisted of around 1800 images which were put on training and 200 images for validation set.

In every training batch during epochs,we also used synthetic dataset used by the authors of CRAFT paper which consists of words placed randomly on some scenic backgrounds. One sample from this dataset was combined with 3 instances of images from the custom dataset generated. This was done so that the model doesn't lose the knowlege it has gained from the previous dataset. So the network saw around 2500 images in each epoch.

We also introduced a number of augmentations to replicate the real world data. Some of the augmenatations used were functional (sin) deformations, shear tranform, introducing random lines, random brightness and noise, random rotation and 3D perspective shifts. Considerable amount of time was spent to tune the hyperparameters used in augmentations since some of the augmentations involve distortion of character regions and hence parallely affect the region scores and affinity scores ground truths.

The training loops of CRAFT being very resource hungry, take very long time to complete.We decided to go with finetuing CRAFT,by borrowing weights from the model which was trained on synthetic data by the authors.We were able to train for around 100 epochs and studied the results on by visually inspecting the outputs on the test set images selected based on edge cases. The model output is a set of 2 images - region and character scores which are super-imposed on top of each othe, as a result the different character regions of a word get connected with affininty scores in between and eventuall this is passed through same post-processing techniques used for ARU Net to get the exact bounding boxes around the words.

Experimental Setup

For training ARU Net, we used a machine with 8GB of dedicated GPU memory and 32GB of RAM. The machine had 12 CPU cores, helping with faster augmentations on the go. The GPU used was of type was NVIDIA GeForce GTX 1070. With this device harwarwe configuration, the training took around 6-7 hours for 100 epochs.We used tensorflow-gpu for model training purposes, open-cv and scikit-learn for augmentations.

For training CRAFT in supervised mode along with augmentations, we procured a virtual machine on azure with 16GB of dedicated GPU memory.The GPU used was Nvidia Tesla T4.The machine had 28GB of RAM and 4 physical CPUs.Even with this high end training setup, an epoch took around 90 minutes to complete.The experiments were run inside gpu optimised ptorch containers provided by Nvidia. We used pytorch for model training with custom dataloaders and used Open-CV for most of the post processing jobs.

We trained ARU Net on historical data, as done by the authors in the paper as well as on custom document data generated by us.After training ARU Net for 100 epochs just on the historical data with no augmentation, we get a base model which performs good on simple documents without any augmentation but misses out the lines in the documents which are randomly oriented and have noise.After 100 epochs, the average loss came to be around 0.05438. This was done to reach to the level of the baseline model provided by the authors. Then after adding augmentations and training on the mixed set of data from IAM, some proprietary images and hostoric documents, it bacame robust on all sorts of images agnostic of noise,distortions etc. After 200 epochs, the average validation loss came to be around 0.07226.Further training of the model, decreases the loss but the effect on the baselines produced is not so considerable in terms of final post-processed outputs we acheive and it saturates after around 110 epochs, the model with best results is also obtained on 110th epoch.The graphs attached here are for training on custom data.

Although the ARU Net can work on images of any size, we found out that keeping the size fixed (768x768px in our case) gives better results.We also trained just U-Net, R-UNet(with residual connections) and AR-UNet(with residual connections and Attention) separately and found out that ARU Net works best to capture better representations and to reach at better baselines.

As far as inferennce results are concerned, the trained model does a descent job on prediction of baselines in most of the cases of machine print but the post processing results using connected component analysis are not accurate in all the cases, we get two boxes for a word/line sometimes and sometimes the word boxes spread to the line below, leading to inaccurate character recognitions in the subsequent steps.Some of the results are attached here with.For handwritten documents which already have lines under the text(lines present in notebook), the baseline predictions are not good enough in those cases.

ARU Net: Graphs

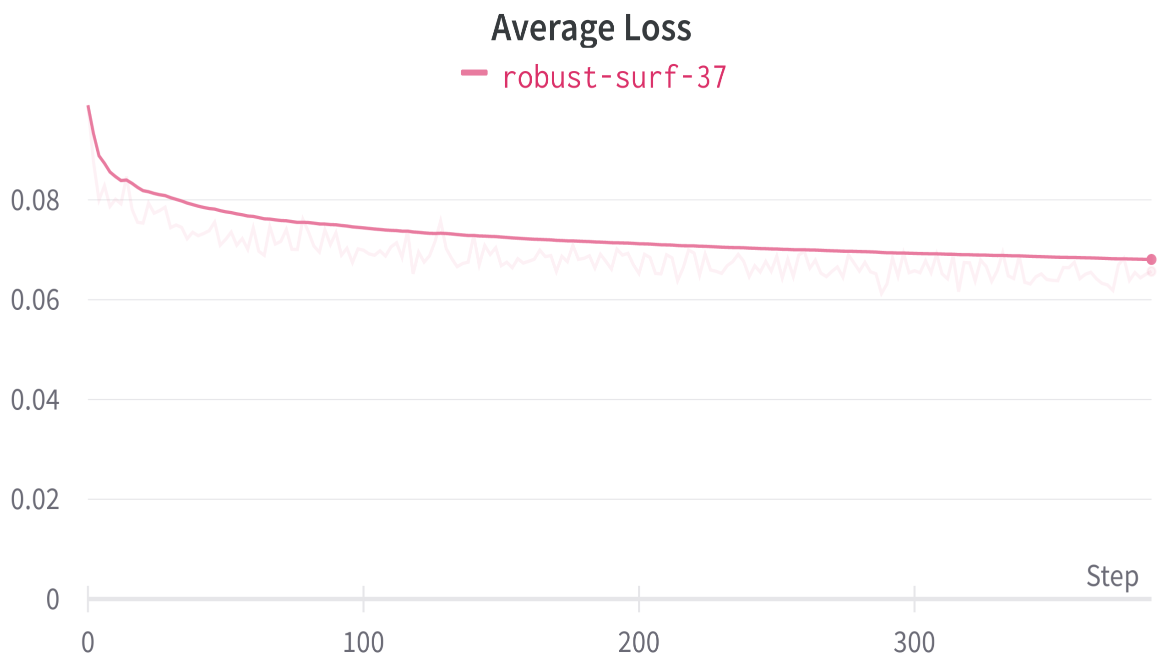


Figure 5.1: Average Training Loss: ARU Net

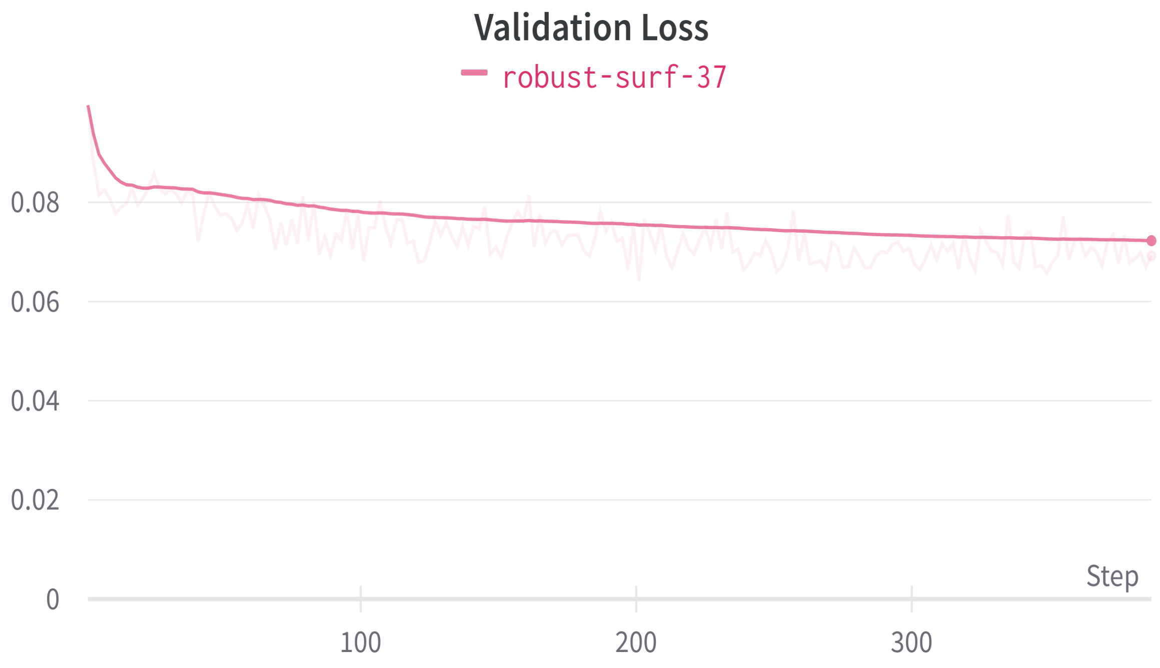
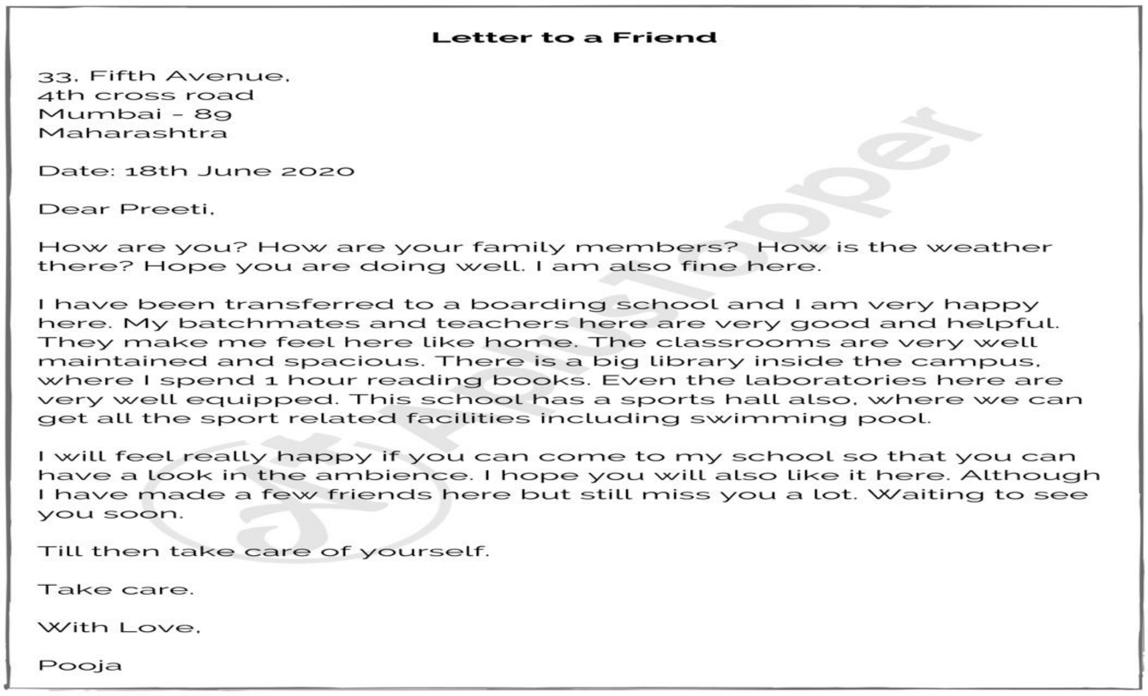


Figure 5.2: Validation Loss: ARU Net



Letter to a Friend

33. Fifth Avenue.
4th cross road
Mumbai - 89
Maharashtra

Date: 18th June 2020

Dear Preeti,

How are you? How are your family members? How is the weather there? Hope you are doing well. I am also fine here.

I have been transferred to a boarding school and I am very happy here. My batchmates and teachers here are very good and helpful. They make me feel here like home. The classrooms are very well maintained and spacious. There is a big library inside the campus, where I spend 1 hour reading books. Even the laboratories here are very well equipped. This school has a sports hall also, where we can get all the sport related facilities including swimming pool.

I will feel really happy if you can come to my school so that you can have a look in the ambience. I hope you will also like it here. Although I have made a few friends here but still miss you a lot. Waiting to see you soon.

Till then take care of yourself.

Take care.

With Love.

Pooja

Figure 5.3: Input: ARU Net

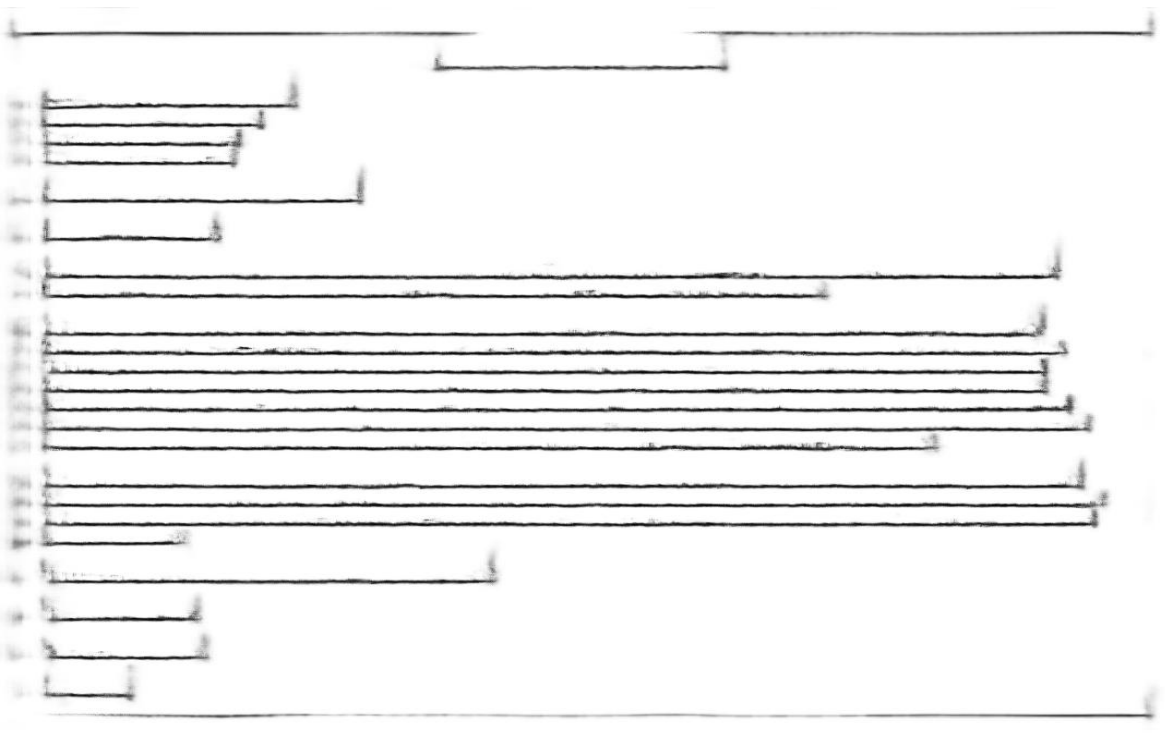
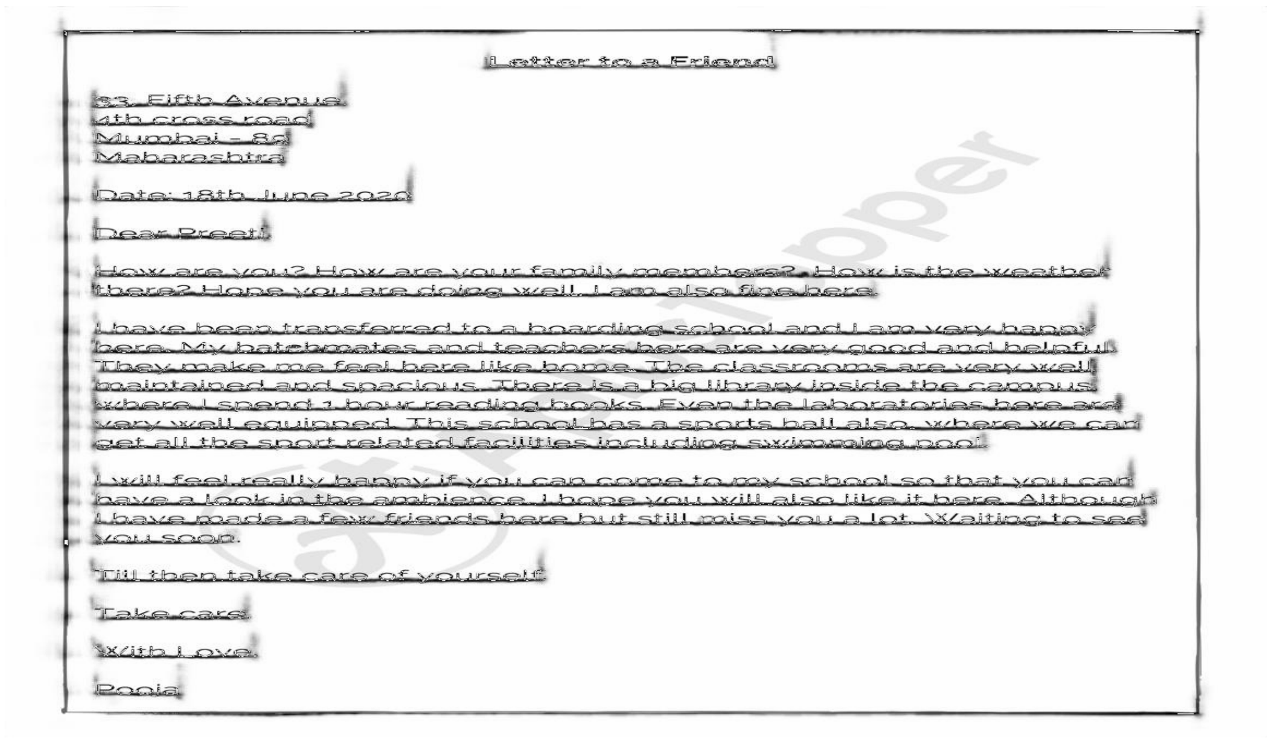


Figure 5.4: Output: ARU Net



Letter to a Friend

33, Fifth Avenue,
4th cross road
Mumbai - 89
Maharashtra

Date: 18th June 2020

Dear Preeti,

How are you? How are your family members? How is the weather
there? Hope you are doing well. I am also fine here.

I have been transferred to a boarding school and I am very happy
here. My batchmates and teachers here are very good and helpful.
They make me feel here like home. The classrooms are very well
maintained and spacious. There is a big library inside the campus
where I spend 1 hour reading books. Even the laboratories here are
very well equipped. This school has a sports hall also, where we can
get all the sport related facilities including swimming pool.

I will feel really happy if you can come to my school so that you can
have a look in the ambience. I hope you will also like it here. Although
I have made a few friends here but still miss you a lot. Waiting to see
you soon.

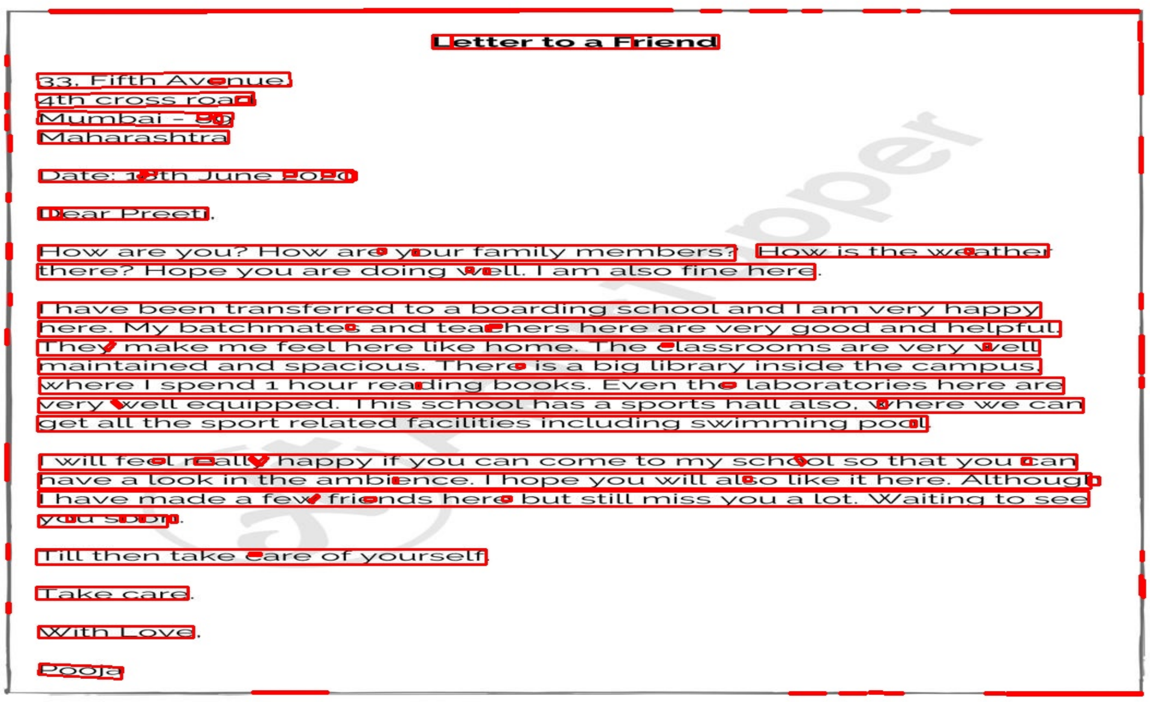
Till then take care of yourself.

Take care.

With Love,

Pooja

Figure 5.5: Intermediate Result: ARU Net



Letter to a Friend

33, Fifth Avenue,
4th cross road
Mumbai - 89
Maharashtra

Date: 18th June 2020

Dear Preeti,

How are you? How are your family members? How is the weather
there? Hope you are doing well. I am also fine here.

I have been transferred to a boarding school and I am very happy
here. My batchmates and teachers here are very good and helpful.
They make me feel here like home. The classrooms are very well
maintained and spacious. There is a big library inside the campus
where I spend 1 hour reading books. Even the laboratories here are
very well equipped. This school has a sports hall also, where we can
get all the sport related facilities including swimming pool.

I will feel really happy if you can come to my school so that you can
have a look in the ambience. I hope you will also like it here. Although
I have made a few friends here but still miss you a lot. Waiting to see
you soon.

Till then take care of yourself.

Take care.

With Love,

Pooja

Figure 5.6: Bounding boxes: ARU Net

Shifting discussion to CRAFT network, the baseline model we started fine-tuning with, was trained on just syhthetic data.Synthetic data was very different than document data where we have hundreds of words in a page as compared to just a few words in syntheic data.Also the size of the words is comparitively small as compared to synthetic dataset.The baseline model was not able to perform good on augmented docuemnts in which text is at random angles.In all, the performance on the document data was not so great.We trained our model for around 100 epochs. An epoch takes around 90-100 minutes to complete and we analysed the results after every epoch. After certain epochs, the results start to degrade rather than improving,this can be due to the reason that the character level bounding boxes in the training data were not accurate and we were doing fully supervised learning of CRAFT.

In the first run, we didn't combine the input data with syhthetic data as used by the authors and as a result, the network started to forget the past learned information and performed bad even on the examples the baseline model performed good, although the results on the handwritten documents improved a bit.

In second phase, in every batch, we combined our custom training data with synthetic data used for training the baseline model we started with,used in the ratio of 3:1 in every batch. The results of training the network with this change helped to stay good on the cases where baseline performs good as well as get better at handwritten text. There are still some issues in recognising the text in the handwritten documents in which ruling lines are present in the paper, for which the model fails to recognise the words.But the results of recognising the words on plain paper are quite good.Certain issues also exist in machine print documents where the CRAFT model fails to recognise small characters like comma,hyphens,points,single characters and it becomes extremely critical in certain business documents.The sample result images are attached here with.

We also tuned the hyperparametes required for connected component analysis and obtaining words from region and affinity scores - link threshold, region threshold and text threshold. Region threshold filters out the pixels regions below certain value and link threshold does that for affinity scores. After combining region and affinity scores, we do connected component analysis and find a large number of components and then we use text threshold to select the appropirate components for finding the minimum area rectangles.For our use case, we selected text threshold as 0.3, link threshold as 0.3 and region threshold as 0.4.

We have deployed the best epoch model which feeds in Word OCR model with the bounding boxes of all the words present in the page.The API is being used internally for testing.

CRAFT: Graphs

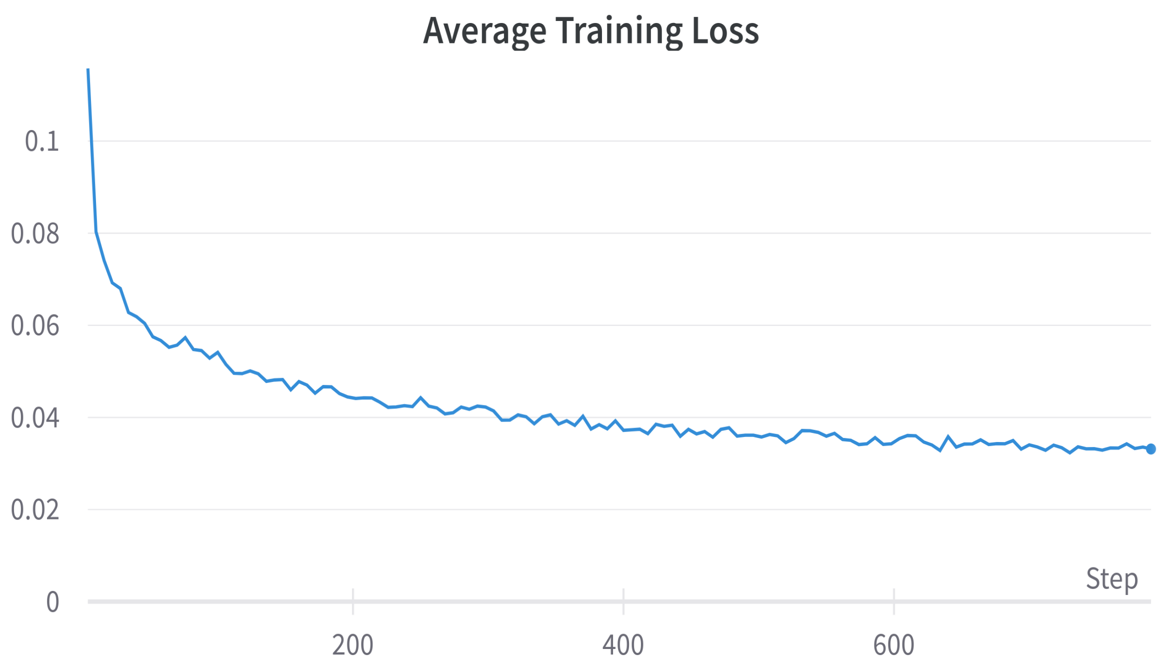


Figure 5.7: Average Training Loss: CRAFT

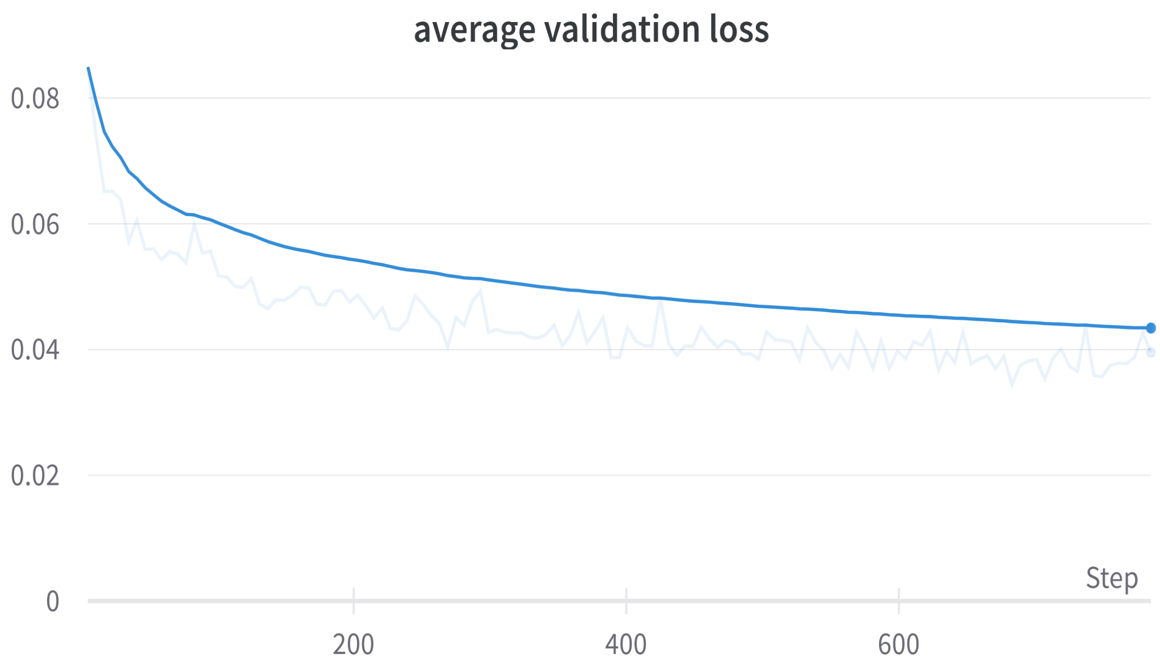


Figure 5.8: Validation Loss: CRAFT

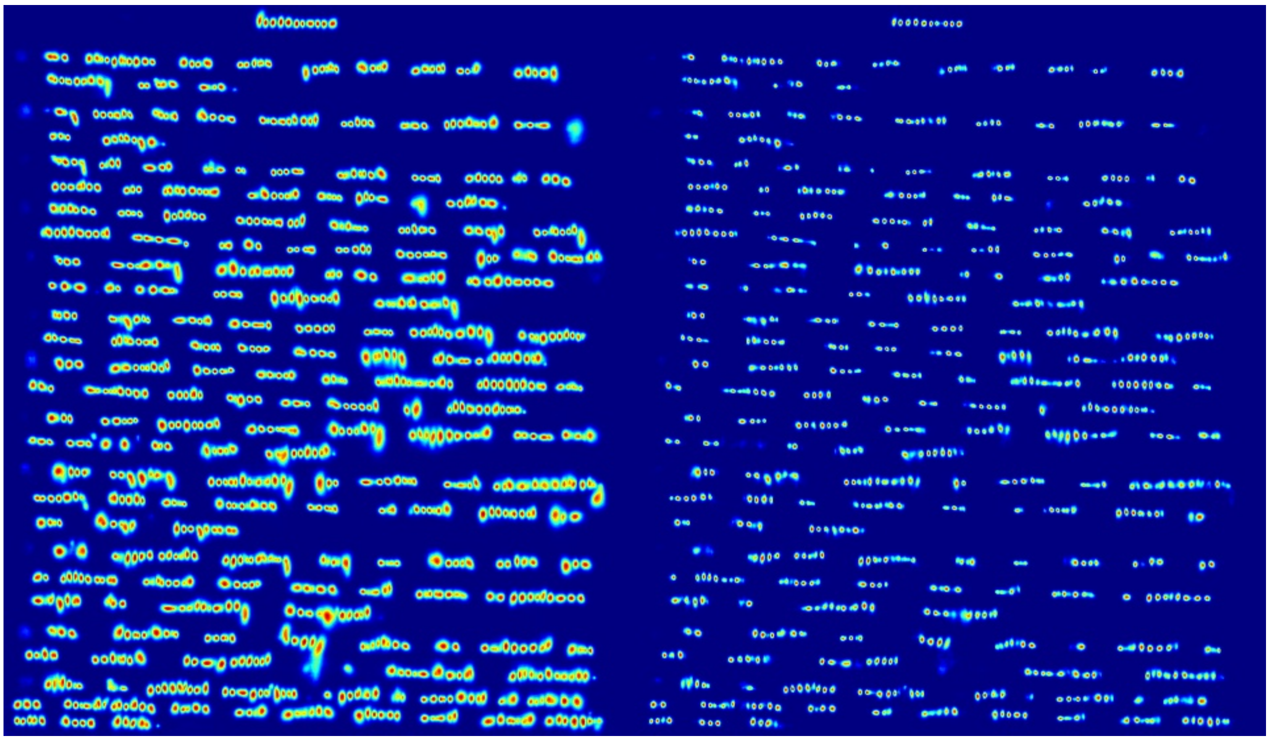
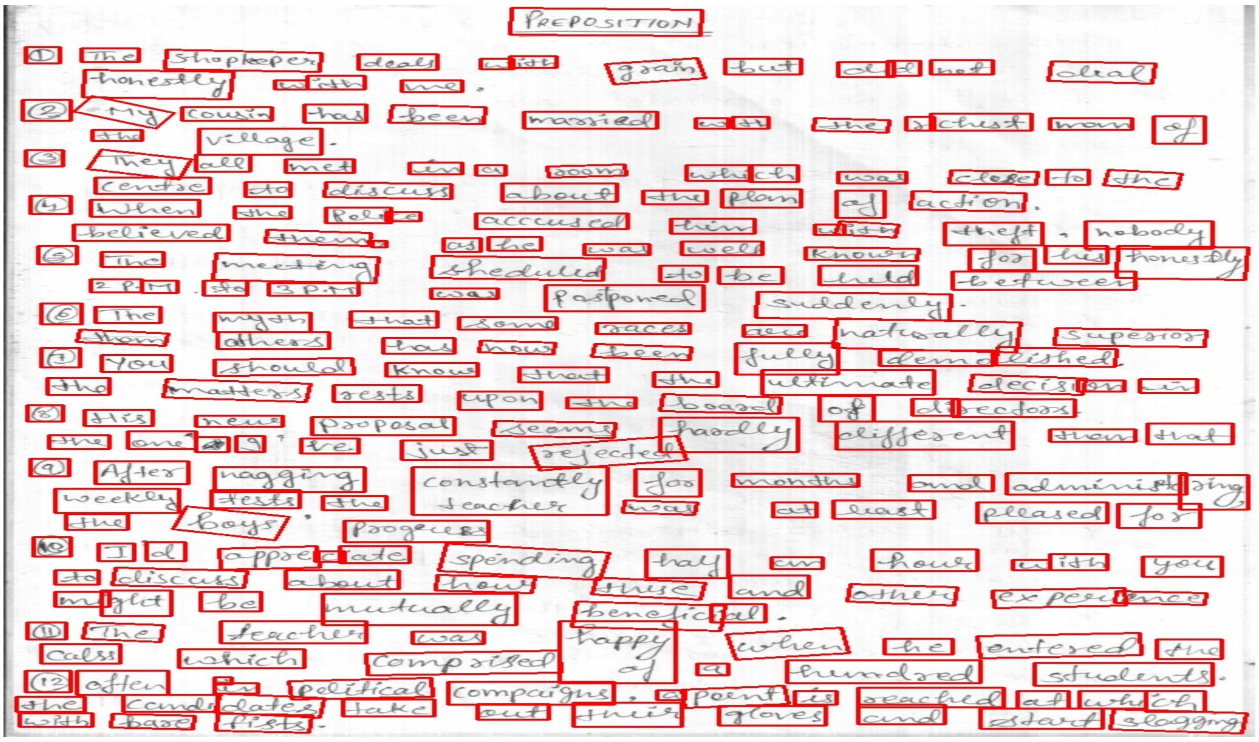


Figure 5.9: Masked output: CRAFT



PREPOSITION

1) The shopkeeper deals with grain but did not deal
honestly with me.

2) My cousin has been married with the richest man of
the village.

3) They all met in a room which was close to the
centre to discuss about the plan of action.

4) When the police accused him with theft, nobody
believed them.

5) The meeting sheduled to be held for his honestly
2 P.M to 3 P.M was postponed suddenly.

6) The myth that some races are naturally superior
than others has now been fully demolished.

7) You should know that the ultimate decision in
the matter rests upon the board of directors.

8) His new proposal seems hardly different than that
the one he just rejected.

9) After nagging constantly for months and administering
weekly tests the teacher was at least pleased for
the boys' progress.

10) I'd appreciate spending half an hour with you
to discuss about how these and other experience
might be mutually beneficial.

11) The teacher was happy when the entered the
calss which comprised of a hundred students.

12) After the political compaigns, a point is reached at which
the candidates take out their gloves and start slogging
with bare fists.

Figure 5.10: Bounding Boxes: CRAFT

Conclusion

We started out with the goal of getting better at text recognition, focusing on handwritten text and identified 3 paths - text segmentation, improving image quality and doing character level segmentation and recognition. We started on the path of text segmentation and tried 2 network archtiectures - ARU Net and CRAFT. As discussed in the results section, ARU Net just gives the baselines and we heavily depend on our post processing techniques to reach the text segmentation we want. This approach does a descent job in identifying the lines but there are certain cases, still not addressed, regarding mis predictions in case of complex handwritten data and irregular number of boxes for a line in case of machine print images. For CRAFT, using supervised learning, we weren't able to reach good results since its difficult to gather character level bounding box data for handwritten documents and the dataset we used, also had several inaccurate character level bounding boxes. We got better at recognising the handwritten text while continuing doing good on machine print document,but the model is not enough good as of now to be production ready.

Ongoing Experiments and future scope: We are exploring the path of semi-supervised learning to get better at segmentation using CRAFT where in we provide the word level bounding boxes and let the the network figure out the character level regions using the interim model trained on synthetic data. Several experiments are also being done for training the model in shorter time, using multiple dataloaders, parallel processing and maximising the GPU usage. We are also planning to work on improving the handwritten OCR model so as to get better at recognising handwriting after segmentation. Some optimisations in the post processing techniques are also underway to get better text recognition results. We have also planned certain changes to original CRAFT implementation in terms of changing the network architecture to do transfer learning to get to the kind of model outputs that help us more in post processing. All such experiments are being worked upon.

Bibliography

[1] Iuliu Konya Marcin Namysl. Efficient, lexicon-free ocr using deep learning. `https://arxiv.org/abs/1906.01969`, 2019.

[2] Thomas Brox Olaf Ronneberger, Philipp Fischer. U-net: Convolutional networks for biomedical image segmentation. `https://arxiv.org/abs/1505.04597`, 2015.

[3] M. Diem F. Kleber S. Fiel T. Grüning, R. Labahn. Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents. `http://arxiv.org/abs/1705.03311`, 2017.

[4] Tobias Strauß Johannes Michael Roger Labahn Tobias Grüning, Gundram Leifert. Deep residual learning for image recognition. `https://arxiv.org/abs/1512.03385`, 2015.

[5] Tobias Strauß Johannes Michael Roger Labahn Tobias Grüning, Gundram Leifert. Aru-net: A two-stage method for text line detection in historical documents. `https://arxiv.org/pdf/1802.03345.pdf`, 2019.

[6] Dongyoon Han Sangdoo Yun Hwalsuk Lee Youngmin Baek, Bado Lee. Character region awareness for text detection. `https://arxiv.org/abs/1904.01941`, 2019.