Improved Image Pre-processing for Better OCR Performance(CSE-402 at 22-02-2021)

Gulam Kibria Chowdhury(170103020033 – gkpalash101@gmail.com); MD Zahid Mahmud Emon(170103020030 – zmemon046@gmail.com); (Supervisor: Tasnim Zahan);

Overview

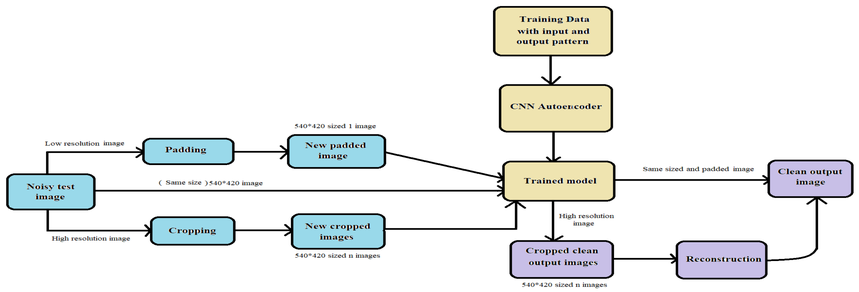
Optical Character Recognition (OCR) has its ups and downs since the beginning due to the preprocessing step for document images. Researches in this topic shows different OCR performance because they all falls behind in different sections due to some restrictions in the methodology. We are using deep learning network and image processing method like (Zero padding, Image cropping and reconstruction) to overcome these restrictions for better OCR performance. We are focusing on variable document images for preprocessing and trying to save all the information so that OCR gives much better performance than existing approaches.

Review of Literature

We have studied different articles about different image preprocessing techniques and the effect of them on OCR improvement. Here are some of the findings-

- Some of the papers focuses on traditional binarization methods but OCR performance didn't improve as much because of the inability to completely denoise the images.
- Some of the papers used Deep learning approach and they had been successful in denoising the images but they had to scale up or down the original image to do so. Though they got clean images, the output images turned out to be of poor quality. Which somewhat affected the OCR results.
- Information loss on image hugely impacts the OCR performance though it's a denoised image.

High-Level-Architecture

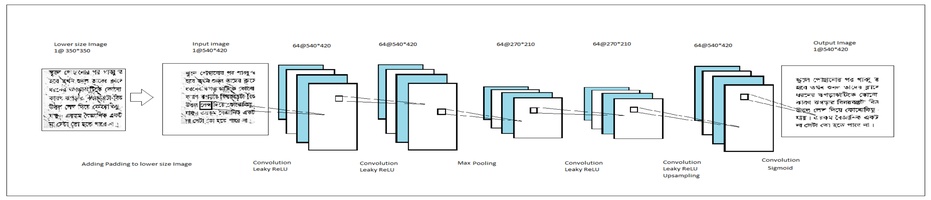


System Architecture

Methodology

We used some preprocessing method that not only denoises the image but also keeps the original image information intact. Zero padding, Image cropping and reconstruction along with denoising deep neural network (CNN Autoencoder) can do that.

1) **Zero Padding:** It is used in convolutional neural network as it refers to the amount of pixels added to an image when it is being processed by the kernel.
2) **Cropping and Reconstruction:** Cropping technique is used for cropping certain portion of an image to use somewhere later and reconstruction of image is a process of adding two or more images into one.
3) **CNN Autoencoder:** A convolutional auto-encoder is a neural network that is trained to reproduce its input image in the output layer. This process is combination of the encoding part and decoding part, where encoder just compresses the image by producing a low dimensional representation of the input image and decoder takes the compressed image and reconstruct the image with only the important data.



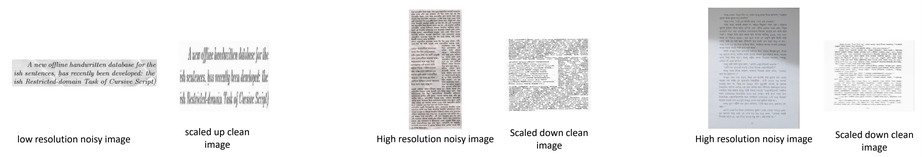
CNN Autoencoder

- We used a 5 layer CNN-Autoencoder model.
- Our model's Encoding part has 3 layers, first 2 layers are 2 convolutional layers and max pooling is used after that.Then 3rd convolution layer.
- Decoding part has 2 layers one convolution layer with upsampling. And another convolution layer which is connected with the output layer.
- Noise pattern is learned in first 3 convolution layer and reconstruction of the original image is done in last 2 layers.
- We used Leaky ReLU in hidden layers and sigmoid activation function on the last layer.

Our Work

We have collected DIBCO (2009-2018) , Kaggle(2015) datasets and created a new dataset of 550 different noise images.

- Previously, high and low resolution images are either been scaled down or scaled up. which distorted the images and produced blurry or stretched out image outputs. And we had to use particular sized images.



low resolution noisy image — scaled up clean image — High resolution noisy image — Scaled down clean image — High resolution noisy image — Scaled down clean image

- We found a way to overcome these problems. For example, we are using Zero padding as a solution for lower size image distortion problem. By adding pixels to images we now don't have to stretch the images and cleaning can be done very easily.

Low resolution noisy image — Noisy padded image — Clean image — Low resolution noisy image — Clean image

- We used another approach for handling large images. We used Image cropping and reconstruction of the cropped images after cleaning. By cropping images into smaller parts cleaning is done without scaling down the large image. Thus output image has all its data intact after cleaning.

Low resolution noisy image — Cropped images — cleaned image — noisy image — cleaned image — High resolution noisy image — Our system cleaned image

Results: We have tested the OCR accuracy for clean images. Accuracy is calculated based on Correct word count. We tested in both Shulikhan and Tesseract OCR engine.

System \ Type of OCR	Shulikhan OCR Accuracy		Tesseract OCR Accuracy	
Size	Low	High	Low	High
Without Pre-processing System	45.2%	40.3%	35.7%	51.1%
Our System	80.9%	98.4%	83.3%	98.7%

Future work

- Resolution enhancement for small images can improve more accuracy for OCR
- More noise can be added to the system though we used most of the noises.

Conclusion

Our approach made a huge jump in accuracy for higher and lower resolution images alongside with trained sized images. Image processing field is much larger and is improving day by day and we think our work will leave a great impression.

Contribution

- We both read more than 10 related papers & implemented some of the research related papers.
- We both contributed on creating our own dataset.
- We both trained different models for our datasets and tested them.
- We both understand our research work and we are ready to answer any question about the research.