Document Analyser Using Deep Learning

Pranali Dhawas[1], Prajwal Kolhe[2], Faizan Khan[3], Lakshya Chauragade[4], Apoorva Dhimole[5]

[1] Assistant Professor, Department Of Artificial Intelligence, G H Raisoni College Of Engineering, Maharashtra, India.

[2345]B.Tech Artificial Intelligence, G.H. Raisoni College of Engineering, Maharashtra, India

---***---

Abstract - *Many companies and massive organizations have numerous documents in bulk and required to keep them in different clusters. In recent years, this job has becoming time consuming as no of document and article has increased. The Analysis of the document is one of the subjective research technique which is used to review by the analyst estimate the idea. With the help of some visual technique we get the most of the layout and text format in extremely in proper output format. With the help of the model we can sort most of the architecture document in large scale. With the help of Layout model and new strategies to interact with various layout in any format in a single model framework. We used our own data collected through out colleagues in university to train it in specific manner that it identify and then classify marksheets and Certificates. Our model use only some of the modern techniques for the visual language task and also for the text images to find the some similar combination and task which make the simple way to detect the interaction between the different stages.*

1. INTRODUCTION

In today's world there is a huge amount of text data, document digitization is a technique which is being used in various fields and domains that deals with huge amounts of archives. Document analyzer focuses on categorizing document on basis of text, image and layout of document, typically documents can be classified on differently on numerous context. While attempting the task of analysing text documents document classification is the important procedure to follow. But while doing document classification we have to deal with some challenges like high variability among the same document or class and low variability between different classes or documents. Previous studies have addressed the structural similarity between classes or documents. Studies has also focused on extracting characteristics to make each class separate. Researchers have developed various deep learning approaches to improve the performance of their document classifiers. In 2014 researcher trained and proposed a simple four layer CNN from starting. Then, we use the Imagenet to improve the learning rate of the network to work in effective way on the model. Most recent research have shown the usage of OCR to extract the particular information and features to with multiple algorithms to classify, NLP was also used to boost the performance of these architecture. Our architecture uses image features, text extraction, and layout of document to overcome pervious defects and drawbacks with significant accuracy.

2. Literature Survey

In Recent time the Nlp and CV became more popular techniques in this area and also make progress in different fields.

- The devlin et al Explained the new language and rendering the model, which is used to see the the representation of bidirectional form the various data by the jointly the condition.

- Barcelona Supercomputing Center – Upgrading the accuracy and edit classification image direct to the parallel system.

- Linkbert - Training the language model by Using Document link.

3. Methodology

Proposed solution in this paper "Document Analyzer" attempts to predict the class of a document by analysing the ssssimage content. To solve this challenge we have addressed three ways – Image classification problem, text classification problem and layout identification. We propose a multi-modal Transformer model to join the document in particular text, layout definition and visual data inside the pretraining stage, the cross-modal learns the interaction in a single framework. The class of document has been predicted according to various features like, the design of the document, header and footer of document, body or the matter of the document which gets extracted by the OCR techniques and how document is being formatted, all of these features help to find the exact class of the given document. But some type of documents also have common features for example government certificate have seal of the govt. and/or logo, which can help classify the documents.

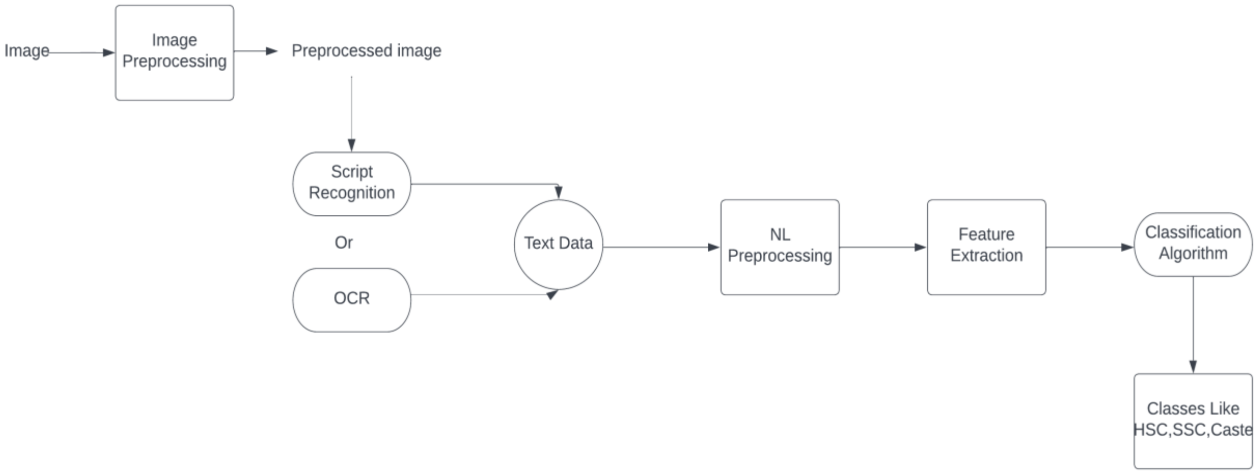


Chart -1: Working Flow

The LayoutLM architecture is divided into 3 section text embedding visual embedding and layoutembedding. In this of the embedding the common work is use to tokenize the ocr and the sequence of the text and giving some segment. The subword is one algorithm used in the Natural language processing. the sentences is used to check the character in the above mention input and to see the most common combination of the various character in the sentence.

Image Processing -This module deals with basic image preprocessing steps. This intends to do basic preprocessing like image rescaling, image resizing and compression, morphological image preprocessing like Erosion andDilation. The output of this module will be similar pre-processed images.

Script Recognition OR OCR - Optical character recognition is a technique to recognize text from a image. Aim of this module is to extract the text this is also called as textraction. This module will give the output as the unlabeled textdata.

Natural Language Preprocessing - The output from the last module which is text data will be pre-processed in this step. Aim of this module is to clean the text data from previous module, we have used several techniques like removal of stop words, stemming, lemmatization, this will further help in feature extraction.

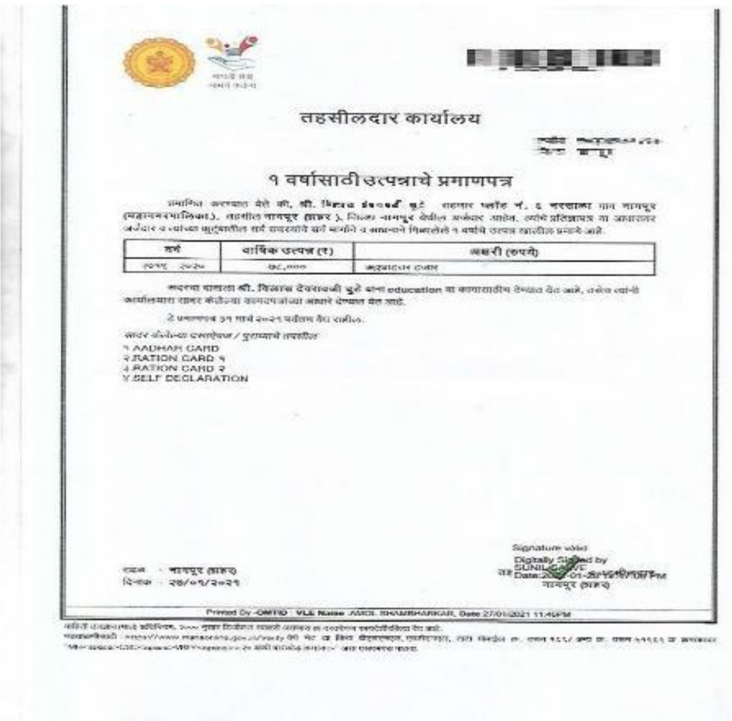
Feature Extraction -Feature extraction is the process in which we try to gain the features from the raw text data intonumerical format.

Encoding -Encoding is also called as vectorization. This process deals with converting the text data into numericalvectors. Converting text data into high dimensional vector format. We can not feed the text data to the machine learning model so we convert the text data into numerical format.

Classification -This module will help to classify the document on the basis of extracted features. We will use classification algorithms like NB classifier, Logistic Regression, Random Forest, etc. this will give theoutput class of the document by assigning one of the class from our available classes.

4. Dataset

In order to train and evaluate document classifier, we have collected above mentioned documents from thestudents from different backgrounds and departments of G. H. Raisoni College of Engineering, Nagpur, Maharashtra. Each class of documents is used in intention to classify the correct details from document image while using OCR technique. Documents are collected from different state students. As these documents are from different states – the layouts of the documents are different for each state in India. For the Document Analyzer we have used various classes to classify the document such as SSC Marksheet, HSC Marksheet, Cast Validity and Income Certificate.



[redacted]

तहसीलदार कार्यालय

[illegible]

१ वर्षासाठी उत्पन्नाचे प्रमाणपत्र

[illegible] श्री. [illegible] नागपूर (महानगरपालिका), तहसील नागपूर (शहर), जिल्हा नागपूर [illegible]

वर्ष	वार्षिक उत्पन्न (रु)	अक्षरी (रुपये)
२०१९ - २०२०	[illegible]	[illegible]

[illegible] श्री. विलास देवरावजी धुरे [illegible] education [illegible]

हे प्रमाणपत्र ३१ मार्च २०२१ पर्यंत वैध राहील.

[illegible]
१ AADHAR CARD
२ RATION CARD १
३ RATION CARD २
४ SELF DECLARATION

ठिकाण : नागपूर (शहर)
दिनांक : २७/०१/२०२१

Signature valid
Digitally Signed by
SUNIL [illegible]
Date:20[illegible]-01-[illegible] PM
नागपूर (शहर)

Printed By -OMTID : VLE Name :AMOL BHAMBHARKAR, Date 27/01/2021 11:40PM

[illegible]

Fig-1: Sample Input Image.

5. Traning Accuracy

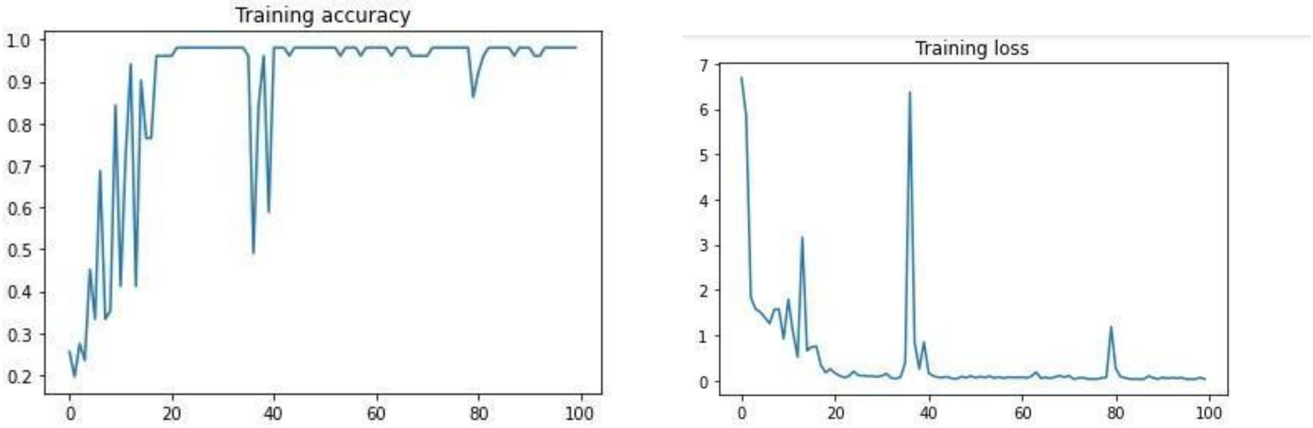


Fig-2: Traning Accuracy And Traning Loss of the Model.

6. Software and Hardware

1.Category: Machine Learning, Deep Learning.

2.Programming Language: Python.

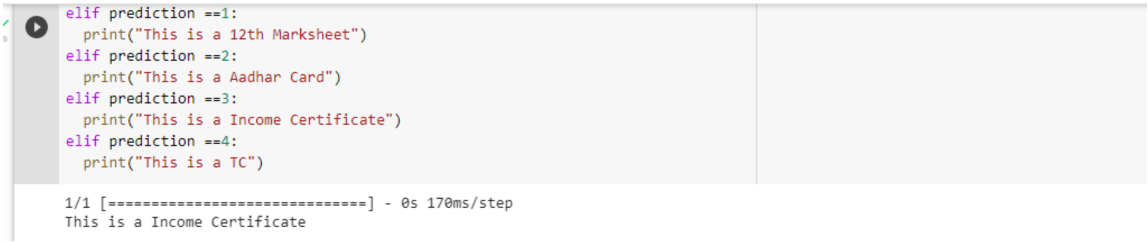
3.Tools & Libraries: CNN, CUDA.

4.IDE: Jupyter, Google Colab.

5.Prerequisites: Python, Machine Learning, Deep Learning, Neural Network

7.Output

As we have uploaded document as shown in the above input phase for the prediction of the document type. We got a desire prediction as shown in below snapshot.



```
elif prediction ==1:
  print("This is a 12th Marksheet")
elif prediction ==2:
  print("This is a Aadhar Card")
elif prediction ==3:
  print("This is a Income Certificate")
elif prediction ==4:
  print("This is a TC")

1/1 [==============================] - 0s 170ms/step
This is a Income Certificate
```

Fig-3: The Output of the Model

8. CONCLUSIONS

We are able to Analyze Document using Pre-Data Training to the model. The model is able to provide a specific Document and information efficient summary. We gave a multi-model pretraining way for visual document understanding tasks.

The Document analyze the categorizing document on basis of the text, image and layout of document, typically document can be classified on different numerous context also. We have compared the different types of document, Original as well as photocopy document while training period and detect the output accordingly.

REFERENCES

[1] Adam W. Harley, A. U. (2015). Evaluation of Deep Convolutional Nets for Document ImageClassification and Retrieval. Toronto, Ontario: Ryerson

[2] N. Chen and D. Blostein. A survey of document image classification: Problem statement, classifierarchitecture and performance evaluation. IJDAR, 10(1):1–16, 2007.

[3] K. Collins-Thompson and R. Nickolov. A clustering-based algorithm for automatic documentseparation. In SIGIR, pages 1–8, 2002.K.

[4] Image and Text fusion for UPMC Food-101 using BERT and CNNs Ignazio Gallo, Gianmarco Ria,Nicola Landro, and Riccardo La Grassa

[5] Aggregated Residual Transformations for Deep Neural Networks Saining Xie, Ross Girshick, PiotrDollár, Zhuowen Tu, Kaiming He

[6] LinkBERT Pretraining Language Models with Document Links Michihiro Yasunaga Jure LeskovecPercy Liang Stanford University {myasu,jure,pliang}@cs.stanford.edu

[7] C. P. Bean, "Fine particles, thin films and exchange anisotropy," in Magnetism, vol. III, G. T. Radoand H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[8] Digitization of Soiled Historical Chinese Bamboo Scrolls. In Proceedings of the 13th IAPRInternational

[9] Workshop on Document Analysis Systems, DAS, Vienna, Austria, 24– 27 April 2018; pp. 55–60.[CrossRef]

[10] Mohammed, H.; Marthot-Santaniello, I.; Margner, V. GRK-Papyri: A Dataset of Greek Handwritingon

[11] Papyri for the Task of Writer Identification. In Proceedings of the 2019 International Conference onDocument

[12] Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019; pp. 726–731.[CrossRef]