International Journal of Scientific Research and Engineering Development-– Volume 4 Issue 3, May -June 2021

Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Image to Text Retrieval

Vibhanshi Modi*, Yash Verma**

*(Department of Information Science and Engineering, Rashtriya Vidyalaya College of Engineering,Bengaluru
Email: vibhanshimodi@gmail.com)

**(Department of Information Science and Engineering, Rashtriya Vidyalaya College of Engineering, Bengaluru
Email:yash.verma24071990@gmail.com)

--************************----------------------------------

Abstract:

Today, as the technology is growing faster than ever and data is being generated at ever increasing rates, a large part of it is covered with unstructured data such as images and videos. It is important to utilize this data too. Though images and videos are useful, textual forms of information have wide acceptance. In particular, high quality and exact semantic information conveyed by text is significant in a wide scope of vision-based application situations. Text retrieval is aimed at detecting and extracting the useful information on the form of text present in these images and videos. Text retrieval from images such as scanned documents, literature and books can be done using the technology known as Optical Character Recognition, OCR. While OCR is very efficient in these scenarios, when applied to normal scene pictures, these Optical Character Recognition mechanisms come up short as they are suited and optimized to the generally high contrast, line-based nature of printed records.

In this paper, we present one such method of retrieving text from natural scene images. The model proposed uses Artificial Neural Networks re-casted as Convolutional Neural Networks which perform much better for image processing than simple neural networks. We have used the YOLO object detection model, which is used for text localization after which the correct text is retrieved by a model which uses the Tesseract OCR engine. With only a few tweaks, the system returned appropriate results.

Keywords **—Text Retrieval,Optical Character Recognition,YOLO.**

--************************----------------------------------

I. INTRODUCTION

Many day-to-day tasks involve processing of a huge number of reports, receipts, forms, invoices, bills, contracts, and a lot more bits of unstructured information, and it's imperative to have the option to rapidly comprehend the data contained inside unstructured information, for faster understanding. Luckily, ongoing advances in computer vision permit us to take extraordinary steps in facilitating the weight of record investigation and comprehension. Optical Character Recognition (OCR) is a technique that is used to convert the images which are in the form of some text such as in scanned documents or typed, handwritten pages. With Screen-text-retrieval (STR) a colossal number of paper-based reports, across different dimensions and applications can be digitized into ML and AI content that makes beforehand out of reach information accessible to anybody at a time. Optical Character Recognition (OCR) is being studied and researched in fields such as Artificial Intelligence(AI), Machine Learning (ML) , Natural Language Processing and Computer vision This is because of the way that newer STRs are prepared by giving them test information which is run over

an AI calculation. Profound learning approaches like deep learning/neural networks can be utilized to join the errands of confiningcontent (Text discovery) in a picture alongside understanding what the content is (Text acknowledgment).
Diversity in fonts, and style of writing make this issue of OCR a challenge to solve. The data is usually in an unstructured format and there is no definite framework of the text. There are a variety of alignments, orientations, fonts, etc depending upon the context and background. Thus, retrieval of this data is an emerging research topic which has witnessed a rapid evolution and advancements in the past few years.

A. Purpose

The purpose of this project is to present the power of Image to Text Retrieval that can solve major real-world applications that drive business impact and research. It can help in the automation of tasks that are time-consuming such as data entry for credit cards, business cards, etc. This can help one extract text from pictures of documents, which can increase accessibility and translate documents resulting in significant reachability. Applications can utilize this technology to improve image-based searches, aided reading for visually impaired, robot navigation, and industrial automation. Thus, there is a major need for an efficient system for the same.

B. Motivation

A lot of time and effort is wasted on a large scale in extracting data from the image. The text extracted has a lot of information that, when put to good use, can produce significant results. Thus, automating this process of retrieval can result in a great deal of optimization of resources and add to the efficiency of the overall process. The wide range of applications discussed above, can be implemented smoothly and result in positive business impact. It can provide sufficient data for further research and algorithm development.

II. LITERATURE SURVEY

The following papers were surveyed to get an understanding of the domain and the existing solutions:

[1] presents comprehension of what CNNs are. CNNs are derived from ANNs keeping a special design in view for the image recognition process. The paper features the advantage of CNNs over other ANNs with the capacity to effectively catch the Spatial and Temporal conditions in a picture.

[2] presented a comparative study of two main single stage object detector CNN models Single Shot Detector and You Only Look Once. SSD model had the advantage over YOLO in that it was able to eliminate most of the False Positive cases which made it more robust. On the other hand, YOLO proved to provide higher accuracy and better detection results as it was able to localize a large number of True Positive cases.

[3] “Scene text detection and recognition: recent advances and future trends” was aimed toward following the new advances in scene text detection and recognition. It featured three strategies for text identification:text-based methods, component-based methods and hybrid methods. It likewise gave an outline of the number of text detection and discovery procedures currently in use with their qualities and shortcomings.

[4] follows a query driven search on pictures and spots the characters of the expressions of a vocabulary in the image database. It at that point processes a score describing the presence of characters of a vocabulary word in each picture. The pictures are then positioned dependent on these scores.

[5] talks about an end-to-end system for text detection – localising and recognising text in common scene pictures.
The segmentation was done by a CNN that regressed the correct coordinates of the bounding boxes. [5]

[6] presents a way to deal with scene text recovery in a single shot. The proposed mechanism depends on YOLO (You Only Look Once), an object detection algorithm which was re-given a role as a PHOC (Pyramidal Histogram Of Characters). This CNN predicts simultaneously the bounding box location, an objectness score and a PHOC of the word in each bounding box. This methodology brought about preferable outcomes over the currently used algorithms as it was performing the process in a single shot.

III. METHODOLOGY

The stepwise procedure we followed is mentioned below.

A.Pre-Processing

First, the image from which the text is to be retrieved is pre-processed. The images found in natural scenes contain a lot of noise. Image pre-processing deals with the removal of this noise. It handles the different lighting in the image, and improves the background of the image for image detection. Albeit pre-handling basically led to improve the picture, as inaccurately applying such channels can represent a danger to the legitimacy of the information gathered. The image below shows the result of pre-processing. It can clearly be seen that there is a significant difference which can result in a better end outcome.

Fig-1 Image Pre-processing

B. Text Detection

Text can be largely classified into two categories - Structured and Unstructured. Structured texts are texts that are present in a standard pattern. This includes text with standard background, proper spacing between each character as well as between each line, standard font families, and a distinctly visible. Unstructured texts on the other hand are text that appear at random with no proper structure in a natural scene. This includes text with no regular font, size, spacing and with a multiplex background making it hard for a detection. Text Detection deals with the detection of unstructured texts. Text Detection involves algorithms that try to notice the presence of text present in a picture or a scene by bounding it inside a rectangular box. This box is known as a bounding box. The bounding box is usually generated by segmenting the images into small pieces and the pieces are matched with one another to detect a region with textual presence. Object detection can be approached in two ways- Region based detection and Single shot detection. In a Region based detector, the first goal is to track down every one of the locales which have the object (text) and afterward pass those areas to a classifier, which gives us the areas of the necessary items. In this way, it is a two-venture measure. Single Shot indicators, in any case, foresee both the boundary box and the class simultaneously.

Here, we have used the Single shot approach as it is much faster, and provided decent accuracy. We used the YOLO model which is a single shot detector. The YOLO (You Only Look Once) model is a cutting edge, ongoing object identification network. There are numerous variants of it. YOLOv3 is the latest and the quickest form. YOLOv3 utilizes Darknet-53 as it's component extractor. It has generally 53 convolutional layers, consequently the name 'Darknet-53'. It has progressive 3 × 3 and 1 × 1 convolutional layer and has some alternate route associations. For the purpose of classification, independent logistic classifiers are used with the binary cross-entropy loss function. Input Image is divided into grids of S x S size by the YOLO algorithm. Every grid cell predicts only one object. This results in limitations on the closeness of these objects. The YOLO model predicts X boundary boxes each if which is associated with a box confidence score. It predicts C conditional class probabilities. Every boundary box consists offive components: (x, y, w, h) and a confidence score for each box. The certainty score reflects how likely the crate contains an article (objectness) and how precise is the limit box. The bouncing box width w and tallness h are standardized by the picture width and stature. Consequently, x, y, w and h are all somewhere in the range of 0 and 1. Every cell has 20 restrictive class probabilities. The conditional class probability is the likelihood that the recognized item belongs to a specific class.

The significant idea of YOLO is to fabricate a CNN network to anticipate a (7, 7, 30) tensor. It utilizes a CNN network to lessen the spatial measurement to 7×7 with 1024 yield channels at every area. YOLO plays out a linear regression utilizing two completely associated layers to make 7×7×2 limit box forecasts. So ultimately, with every detected object, it gives the probability as well of correct detection.

After this the Darknet neural network system is used for preparing and testing the data. The system utilizes multi-scale preparing, heaps of information expansion and batch normalization. It is an open source neural network structure written in C and CUDA. It is quick, simple to introduce, and upholds CPU and GPU calculation. Once the data was collected, it was labelled and annotated with appropriate tags.

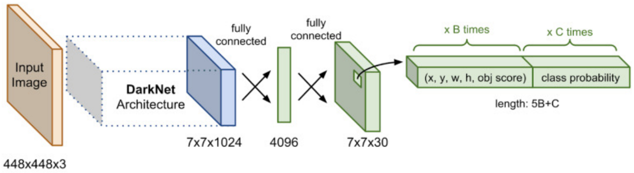

Fig-2

C. Text Recognition

Text Recognition deals with the extraction of the text from the output that comes from the text detection process. Image Recognition mechanisms are based on Convolutional Neural Networks (CNN) which are a special type of Artificial Neural Networks. The output from the text detection step enclosed in a box is used as an input, and feature selection process is applied on it. Convolutional Feature maps are developed using Neural Network that help in feature extractions.

Since we have our custom content locator carried out for text identification, we now move onto the ensuing cycle of Text Recognition. We have used tesseract for the same. The Tesseract OCR engine performed decently with only a few tweaks. The important part now lies in combining the text detection and text recognition. i.e. find the needed region from the image, pass it to tesseract, extract the results of tesseract and store it in a meaningful format. YOLO distinguishes the necessary content districts and harvests them out from the picture. Afterward, we pass those areas individually to tesseract. Tesseract understands them, and we store that data.

So, the overall framework is as represented in the block diagram shown below.

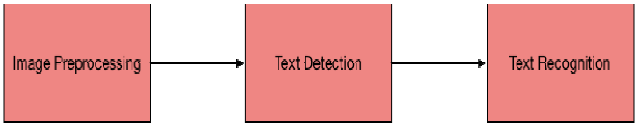

Fig-3 Image to Text Processing

The pretrained weights for the CNN model helped in easy implementation of the network. There is always a trade off between speed and accuracy when choosing between dingle shot and region based detectors. We decided to go ahead with YOLO mainly because of its extremely fast processing speed. There are imitations when it comes to the closeness of objects in the picture.

IV. IMPLEMENTATION DETAILS

The code is tested over the IIIT-R dataset which is reaped from Google and Flickr image search. Inquiry words like coffeehouse, inn, mailing station, school, office was utilized to gather the pictures. Also, question words like sky, building were utilized in Flickr to gather some irregular distractors (pictures not containing text). The dataset contains 10,000 pictures on the whole. The pictures are physically commented on to say if they contain an inquiry word. This dataset consists of pictures that contain a word and for each such word around 100 such pictures are present in the dataset. It contains pictures that contain text in varying fonts, under different lighting conditions, with complex backgrounds, from varying angles, etc. There are some pictures that do not contain any text component as well. This is mainly to enhance the diversity and ensure that the model is trained for noisy data points like these.

The entire system is implemented in the Python programming language. Majority of the functionalities are governed by the use of significant python libraries and packages like OpenCV, a library intended to tackle computer vision issues. OpenCV-Python utilizes Numpy, which is an library for mathematical computing and statistics uses a structure similar to that of MATLAB. OpenCV structures are based on these arrays that come from Numpy library. Tensorflow, has an exhaustive, adaptable flexible system of tools, libraries, and local area assets that performs ML and DL tasks without any problem. TensorFlow gives stable Python and C++ APIs, a lot of which (tf.data, tf.config, tf. slim, and so forth) have been utilized here. Numpy, a library that supports working with large, multi-dimensional arrays and matrices, by providing a large collection of high-level mathematical functions to operate on these arrays. It is intensely used here, as all the image data is converted to arrays which are further processed. Yolov3, to find the best anchor box sizes for the given dataset. Sklearn, a Machine Learning library highlights different arrangement, and bunching calculations including classification, regression and clustering algorithms including SVM, random forests, k-medroid/means and DBSCAN, and is designed to work with the Python numerical and scientific libraries NumPy and SciPy. Here, NearestNeighbours' score and average precision that is printed at the end is calculated using sklearn packages. Scipy, contains packages for optimization, linear algebra, integration, interpolation and many such special functions. Here, the 'io' module of Scipy is used to prepare the queries to perform the image search.

A few preprocessing techniques used for cleaning the images include- de-slant, line expulsion, format investigation and standardization. Albeit pre-handling basically led to improve the picture, inaccurately applying such channels can represent a danger to the legitimacy of the information gathered. The system was tried on different sets of images. The ICDAR dataset was also tried upon but the IIIT R dataset fit more relevant to the situation later.

V. TESTING

The system was tested for various use case scenarios. i.e. for images having text in different fonts, orientations, backgrounds, angles, rotations, sizes, etc. It turned out to perform decently well on all the different types of orientations but for text located very close to each other, the model could detect only either of the bounding boxes. Text in any language other than English could not be detected as well.

VI. RESULT AND ANALYSIS

The proposed system calculates a mean average precision. Average Precision which is defined as the area calculated under the curve of precision recall. Precision is given as the number of True Positives by the sum of True Positives and False Positives indicating the number of predictions that turned out to be correct. A recall on the other hand indicates wellness of the predictions made by the model and is given by the number of True Positives by the sum of True Positives and False Negatives. The mean of all the AP values found for the dataset is mAP. After the execution of the proposed model over the STR dataset the mAP value was calculated to be around 0.6949578.

Here are few key takeaways about Single Shot Detectors:

Single shot detectors have a lovely great edge each seconds (FPS) utilizing lower goal pictures at the expense of exactness. Those papers attempt to demonstrate they can beat the locale based finders' precision. Nonetheless, that is less convincing since higher goal pictures are frequently utilized for such cases. Indeed, single shot and area-based finders are getting a lot of comparable in plan and executions now. However, with some booking, we can say:

- Area based finders like Faster R-CNN exhibit a little exactness advantage if continuous speed isn't required.
- Single shot locators are here for continuous handling. Yet, applications need to check whether it meets their precision necessity.

VII. APPLICATIONS

With OCR, an immense number of research-based archives, across different dimensions and applications can be digitized into ML/AI content, that not just makes collection simpler, yet in addition makes beforehand out of reach information accessible to anybody at a time.

The Banking business, alongside other money area enterprises like protection and protections, is a significant buyer of ITR. The most successful utilization of OCR is to deal with cheques: a manually written check is processed and transformed over into computerized text, which can further be analysed without human inclusion and reduces the bottleneck involved in scaling.

Hardly any businesses produce as much administrative work as the lawful business, thus ITR has numerous applications in this. Reams and reams of testimonies, decisions, filings, proclamations, wills and other authoritative records, particularly the printed ones, can be digitized, put away, database and made accessible utilizing the least complex of OCR per users.

Another industry that deals with ITR on a huge daily basis is the healthcare industry. Reports from different medical tests, prescriptions, tablets, injections, etc can be subjected to retrieval of text printed and thus utilized for further analysis of success and failure based on previous history if available.

VIII. CONCLUSION

In today's world where the amount of data is increasing faster than ever, texts provide a lot of information compared to other types of data. Text is crucial for any analysis as it yields more accurate and faster results attributing to its simplicity to be interpreted by the humans as well as the machine. A growing study is involved in retrieval of text from data sources like images and videos. This project also attempts to employ an algorithm for scene text retrieval. A YOLOv2 object detector is used to detect the text in the images and tesseract is used on the detected text to identify that text more accurately. The results were analysed and it was noticed that the model was able to recognise text with a good average precision for most parts of the dataset. A low average precision was seen for the pictures that contained blurred text or text with styles not seen before. The top 10 results from the analysis were also recorded for further investigation.

IX. FUTURE WORK

Even though the proposed model was able to produce good results for the dataset given, it experienced some trouble detecting the texts, styles and fonts that were not seen before. It also struggled with pictures that were blurry and had complex backgrounds; the model was unable to extract the distinguishable features of text from the background. The future of text retrieval lies in the research of developing the algorithms that can tackle these problems to produce better results. The future work must explore other character representations to extract the text post localisation. The future of text retrieval is promising, since a huge number of images and videos are now posted regularly with the growth of social media platforms. Text retrieval from videos is also an interesting field that must be explored. This can be done by using the current image retrieval algorithms and optimising and fitting them for extracting texts from videos.

REFERENCES

[1] O'Shea, K., Nash, R.: An introduction to convolutional neural networks. CoRR, abs/1511.08458 (2015)

[2] Morera, Á.; Sánchez, Á.; Moreno, A.B.; Sappa, Á.D.; Vélez, J.F. SSD vs. YOLO for Detection of Outdoor Urban Advertising Panels under Multiple Variabilities. *Sensors***2020**, *20*, 4587.

[3] Zhu, Y., Yao, C. & Bai, X. Scene text detection and recognition: recent advances and future trends. *Front. Comput. Sci.* 10, 19–36 (2016)

[4] Anand Mishra, KarteekAlahari, C.V. Jawahar. Image Retrieval using Textual Cues. ICCV - IEEE International Conference on Computer Vision, Dec 2013, Sydney, Australia. pp.3040-3047, ff10.1109/ICCV.2013.378ff. Ffhal-00875100f

[5] Jaderberg, M., Simonyan, K., Vedaldi, A., Zisserman, A.: Reading text in the wild with convolutional neural networks. International Journal of Computer Vision 116(1) (2016) 1–20.

[6] Lluís Gómez, Andrés Mafla, MarçalRusinol, DimosthenisKaratzas, "Single shot scene text retrieval", Proceedings of the European Conference on Computer Vision (ECCV), pp 700-715

[7] Chaitanya R. Kulkarni, Ashwini B. Barbadekar, "Text Detection and Recognition: A Review", International Research Journal of Engineering and Technology (IRJET) Volume: 04 Issue: 06 | June-2017