International Journal of Enhanced Research in Science, Technology & Engineering
ISSN: 2319-7463, Vol. 12 Issue 5, May-2023, Impact Factor: 7.957

Text Detection & Learning Using CNN & RNN

Ronak Thakare[1], Lokesh Kale[2], Kishor Pathak[3], Sonali Bhoite[4]

[1,2](UG) Vishwakarma Institute of Information Technology
[3,4]Assistant Professor, Department of Information Technology, VIIT, Pune.

ABSTRACT

The study of computer vision, including text identification and classification, has been transformed by deep learning. In this paper, we discuss recent developments in deep learning-based text detection and learning. The deep learning-based methods for text recognition, such as You Only Look Once (YOLO) models, faster region-based convolutional neural networks (Faster R-CNN), and single-shot detection (SSD), are first given a general overview. We then go over deep learning-based methods for text identification, such as transformer-based models, attention-based models, and convolutional recurrent neural networks (CRNNs). Our discussion of deep learning-based text detection and recognition uses in a variety of fields, including document analysis, scene text recognition, and video monitoring, comes to a close.

Keywords: OCR, OCR Challenges, OCR Applications, CNN, RNN.

INTRODUCTION

Text detection is a critical component in many computer vision applications such as document analysis, scene recognition, and image retrieval. Traditional methods for text detection relied on handcrafted features and rule-based algorithms. However, these methods often fail to handle the variation in font styles, sizes, orientations, and background clutter. With the advent of deep learning, text detection has witnessed significant progress in recent years.

Deep learning-based text detection models typically use sliding window or region proposal-based approaches. In sliding window-based methods, the model scans the entire image at different scales and aspect ratios to detect text regions. Region proposal-based methods, on the other hand, Create potential text areas using selective searching or similar techniques, and then use a CNN to determine if they are text or not. Once text regions are detected, they can be further processed using text recognition algorithms to recognize and extract the actual text content.

To train a deep learning-based text detection model, a large dataset of annotated images with text regions is required. The model is trained using a supervised learning approach, where the model learns to predict the presence of text in an image based on its input features.

Optical character recognition (OCR), commonly referred to as text recognition, is the process of turning scanned or handwritten text images into editable, searchable digital text. A kind of machine learning called "deep learning" has excelled at text recognition tasks.

Convolutional neural networks (CNNs) are frequently used in deep learning models for text recognition in order to extract features from the input picture and recurrent neural networks (RNNs) in order to recognise the text sequence. A sizable collection of photos with accompanying ground-truth text labels is used to train the model.

By putting a new image into the trained model and receiving the predicted text sequence as output, the model may be used to identify text in fresh photos. Deep learning has applications for text recognition in areas including document digitalization, automated captioning, and handwriting identification.

Text recognition problems may be performed using deep learning in a number of ways. Following are some typical applications of deep learning for text recognition:

OCR, or optical character recognition OCR is the process of reading characters from printed or handwritten text and encoding them into machine-readable text. To identify characters in photos, OCR systems often employ deep learning techniques like convolutional neural networks (CNNs).

NLP stands for natural language processing, which is the study and comprehension of human language. Recurrent neural networks (RNNs) and transformer models, two types of deep learning models, have been used to NLP applications including sentiment analysis, language translation, and text categorization.

Handwriting recognition is the process of identifying handwritten characters and converting them into computer-encoded text. Deep learning models like CNNs and RNNs have been used in handwriting recognition systems.

The act of finding and classifying named entities in text, such as people, groups, and locations, is known as named entity recognition (NER). Deep learning models, like as transformer models, have been used to solve NER issues. Analyzing the layout of a document entails examining its organization, including its headings, paragraphs, and other elements. Deep learning models like CNNs have been used in document layout analysis systems.

LITERATURE REVIEW

Several scholars have studied the different methods for finding text in photographs. A sliding window idea was utilized to assess the distinctive texture contained in the input picture in certain studies that investigated the texture-based strategy for identifying the text contained within the picture. A recognition technique based on the MNIST, NIST, and IAM datasets, normalization, and training the dataset for picture categorization was proposed by Brindha Muthusamy[1].

--input: A photograph, for example, is a single-object image.
--Output: A designation for a class(one or more integers that are mapped to class labels).
validating the datasets, and testing the dataset:

Testing data demonstrates that once a model has been established, it is able to producing accurate predictions. The test data should not be labelled if labels are being used to track the metrics of the model in the training and validation data. Test results are used to verify an ambiguous dataset and make sure the machine learning algorithm was properly trained.

Object detection (recognizing the various kinds or classes of items in a picture by using a boundary to detect their existence).

o Input: A photograph, for instance, is an image featuring one or more objects.
o Output: A class label for each bounding box, as well as one or more bounding boxes (specified by a point, width, and height).

Other researchers concentrated on computer vision applications' usage of sparse-based text identification techniques[2]. It has been put out by Mingyu Wang[2]. These techniques help create edge maps from the picture. When classification has been completed, a second sliding window is utilized to extract the textual patches .

The unique technique for scene text detection put forward by Ashish Kumar Jha[3], Sandeep Kumar, and MVV Prasad Kantipudi[3] combines bidirectional LSTM and deep convolution neural networks. Using the described method, the contour of the image is first determined, and then it is sent to CNN. CNN is used to create the features in a series of orders from the contoured image. The BiLSTM is used to rank the characteristics as of right now. The Bi-LSTM is a helpful tool for taking characteristics out of a word series. As a consequence, they combine the two efficient processes for extracting characteristics from the input image that is based on contours and the input picture, which speeds up recognition and enhances the strategy compared to existing methods.

According to research by John Clark[4], Randy Eugene[4], Dylan Bryan[4] and Bruce Albert[4], text is the visible and tangible carrier of human culture, and the discovery and recognition of text strengthens the link between vision and content comprehension.

BACKWARD LITERATURE

Table 1: Backward Literature

Reference no	Authors	Methodology	Application	Limitation	Future scope
1	Brindha Muthusamy, Kousalya K	method follows: preliminary steps for text detection and recognition for that it use Datasets, image pre-processing(data integration, data reduction, data transformation), Normalization for the organizing data in the database, and then text classification using Long-Short Term Memory Networks(LSTMs)	Document digitization: Text detection and OCR are commonly used for digitizing physical documents, such as books, newspapers, and historical records, into machine-readable formats.	achieves good results for tilted and bent texts. However, there are still large noises and relatively deformed shapes in natural scenes	Optical Character Recognition will play a vital role to find a way to digitize the words and numbers in physically written text and characters in different languages.
2	Mingyu Wang	A text recognition model combining two dimensional attention mechanism and CTC is designed, combined with the two related work, the framework is integrated and extended to an end-to-end recognition system.	Image and video search: Text detection and recognition are used in search engines to identify and index text that appears in images and videos. This allows users to search for visual content using keywords and phrases.	SegLink fails to link the characters with large character spacing.	Text detection and recognition are used in search engines to identify and index text that appears in images and videos.
3	Sandeep Kumar ,Ashish Kumar Jha , and MVV Prasad Kantipudi ,	The image's contour is recognized, and this information is sent into CNN, which creates a sequence of ordered of the features that are then coded using BiLSTM. Bi-LSTM is used to extract the features from a series of words. In order to speed up the recognition process, this research combines two effective methods for extracting features from images and contour-based images.	Automated translation: Text detection and machine translation are used to automatically translate text from one language to another. This technology is being used in various industries, such as e-commerce, travel, and healthcare.	If background is complicated, hazy, and lit differently, these algorithms cannot produce superior outcomes. When the techniques are used with the actual dataset, the computing cost is quite high.	NLP is an area of computer science that deals with the interaction between computers and humans in natural language. Deep learning techniques can be used to improve the accuracy of NLP algorithms and enable computers to understand and generate human language better.

4	Dylan Bryan, Randy Eugene, John Clark, Bruce Albert	The use of deep learning-based models, which have significantly changed how academics approach challenges and broadened the field of study, is the key area of concentration. In addition, the paper discusses current techniques in a top-down fashion and organizes them into four systems: text detection, recognition, end-to-end, and auxiliary approaches. Each category's most recent techniques are examined from various angles. The paragraph also emphasizes the value of text recognition in a number of contexts, including text-to-speech devices, sign recognition on the road, and note digitalization.	Autonomous vehicles: Text detection and recognition are being used in autonomous vehicles to identify road signs and traffic signals. This helps the vehicles navigate the roads and make informed decisions.	Although the majority of publications state that they train their models to recognise case-sensitively and to incorporate punctuation, they may simply output numbers and situation-insensitive characters throughout the assessment phase.	Improved accuracy and speed,Improved image preprocessing,Integration with other technologies,Improved image preprocessing

PROBLEM STATEMENT

The main objective of this Problem Statement is to detect the text from image and drawing a bounding box to that text and then recognize that text from bounding box.

Challenges in text detection:

(a)Poor quality text due to various climatic conditions
(b)Multilingual text.
(c) Multi-coloured text
(d) Curved text.
(e) Images with shadow

METHODOLOGY

Text detection and recognition using deep learning is a complex process that involves several steps. Here is a general methodology that can be used for text detection and recognition using deep learning:

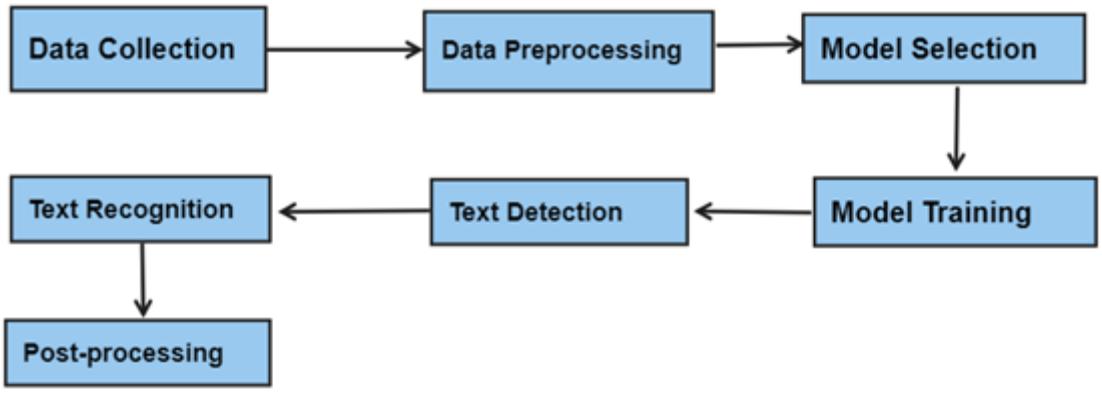

Fig 1 step for text detection

1. **Data collection:** assemble a database of text-filled pictures. To guarantee that the model can identify text in a variety of situations, the data set should include pictures of various sizes, resolutions, and typefaces.
2. **Data preprocessing:** Preprocess the dataset to make sure the pictures are all the same size and style before training the model. Image normalization, such as changing the brightness and contrast, may also be part of this process.

We need to do some preprocessing in order to make it suitable for our model. Both the input picture and the output labels require preprocessing.

- Read the image and create a grayscale version of it.
- Use padding to make each picture 128,32 pixels in size.
- Increase the image's size to (128,32,1) to ensure that it will work with the shape provided for architecture.
- Divide the image's pixel values by 255 to normalise it.

We se the following methods to preprocess the output labels:

- Read the text from the image's name, which contains content that has been written inside the picture.
- Create a function that converts each letter of a single word into a number value, such as "a":0, "b":1, "z":26, etc. If we had the word "abab," for example, our encrypted label would be [0,1,0,1].
- Calculate the maximum word length and enlarge each output label to match the maximum length. To make it consistent with our RNN architecture's output form, this is done.

3. Model selection: A suitable deep learning framework should be chosen for text detection and classification. CNNs and RNN are two popular designs for text detection.

- **CNN** (Convolutional Neural Network)**:** which is a class of deep learning algorithm commonly used in work involving image and video identification. CNNs use a series of convolutional layers, pooling layers, and completely linked layers to autonomously learn and derive features from raw input data, such as images.

  Pooling layers down sample the feature maps to reduce the density of the data and avoid over fitting, while convolutional layers apply filters to the input picture to extract significant features like edges, textures, and patterns. Based on the extracted characteristics, the original picture is classified using fully linked layers.

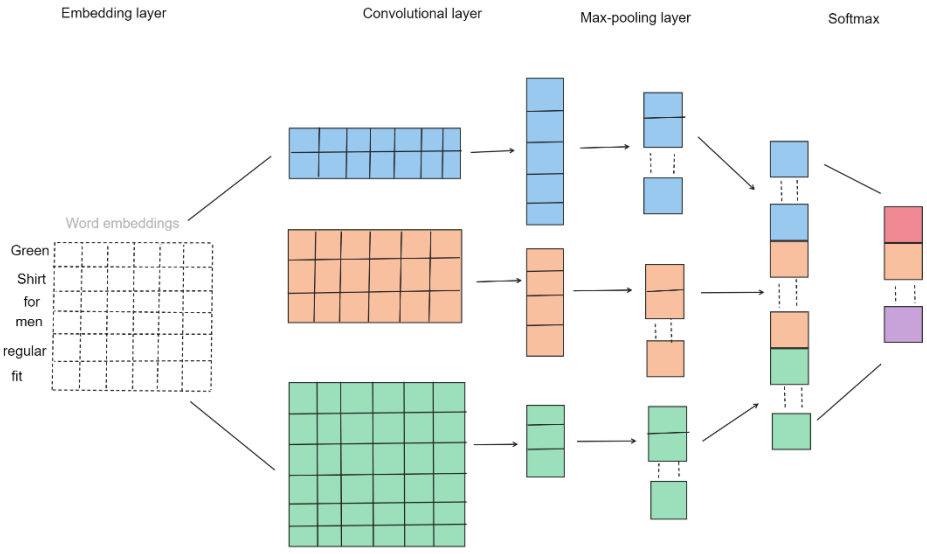

Fig 2: CNN architecture for text classification

- **RNN** (Recurrent Neural Network): which is a class of deep learning model frequently employed in jobs involving linear data analysis and natural language processing (NLP). By utilising feedback

links, which enable the network to keep track of prior inputs, RNNs are created to handle sequential data.

The main strength of RNNs is their capacity to handle input sequences of various lengths, which makes them ideal for tasks like mood analysis, language translation, and voice recognition. RNNs accomplish this by capturing relationships between inputs by using a hidden state that is changed at each time step.

The LSTM network is a well-liked RNN variation that features extra gates to regulate the flow of data into and out of the concealed state, enabling the network to preferentially recall or forget information.

4. Model training: On the pre-processed dataset, train the chosen model. In order to enhance the efficacy of the model, this stage entails choosing the proper the hyper parameters such as the rate of learning and batch size.

5. Text detection: Utilize the trained algorithm to find text in fresh pictures. Object recognition methods like sliding windows or region-based approaches may be used in this stage.

6. Text recognition: Use a model that has been trained to identify the text in the picture once text has been identified. To decode the observed text in this phase, sequence-to-sequence models like RNNs or transformer models may be used.

7. Post-processing: Post-process the text that has been recognized in order to increase accuracy and fix any mistakes. This stage might entail using model languages to fix contextual or spelling problems.

CONCLUSION

Text detection and identification methods based on deep learning have demonstrated amazing success in a variety of text detection and recognition challenges. There are numerous uses for these methods in various industries, such as document analysis, scene text identification, and video monitoring. The extraction of text from twisted or deformed symbols as well as the detection of text in low-resolution pictures are still issues that need to be resolved. To resolve these issues, future study should concentrate on creating more reliable and effective deep learning-based text identification and recognition methods.

REFERENCES

[1]. **Brindha Muthusamy, Kousalya K .”** Deep Learning in Text Recognition and Text Detection: A Review”, Dept. of Computer Science and Engineering, Kongu Engineering College, Erode. published by International Research Journal of Engineering and Technology (IRJET) Volume: 09 Issue: 08 , Aug 2022.

[2]. **Mingyu Wang.** “Deep learning based text detection and recognition in natural scenes”, 2021.

[3]. **MVV Prasad Kantipudi, Sandeep Kumar, and Ashish Kumar Jha.** “Scene Text Recognition Based on Bidirectional LSTM and Deep Neural Network”, Department of E&TC, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India, published by Hindawi Computational Intelligence and Neuroscience , 23 November 2021.

[4]. **John Clark, Randy Eugene, Dylan Bryan, Bruce Albert.”**A Study of Text Detection and Recognition”, published by Portland Community College , July 2021.

[5]. **Karez Abdulwahhab Hamad, Mehmet Kaya,”**A Detailed Analysis of Optical Character Recognition Technology ”, published by International Journal of Applied Mathematics, Electronics and Computers, ISSN: 2147-8228 , December 2016.

[6]. **Chuang Yang, Mulin Chen, Yuan Yuan, Senior Member, IEEE, Qi Wang.”**MT: Multi-Perspective Feature Learning Network for Scene Text Detection ”, 12 May 2021.

[7]. **Md. Anwar Hossain & Sadia Afrin.”**Optical Character Recognition based on Template Matching”, published by Global Journal of Computer Science and Technology: C Software & Data Engineering, Volume 19 Issue 2 Version 1.0 Year 2019.

[8]. **M. IKONOMAKIS , S. KOTSIANTIS & V. TAMPAKAS ,** “Text Classification Using Machine Learning Techniques” published by WSEAS TRANSACTIONS on COMPUTERS, Issue 8, Volume 4, August 2005.