International Journal of Education, Modern Management, Applied Science & Social Science (IJEMMASSS)
ISSN : 2581-9925, Impact Factor: 6.882, Volume 05, No. 04(III), October - December, 2023, pp. 177-183

ANALYSING THE EFFECTIVENESS OF CNN-RNN IN RECOGNIZING SCENE TEXT

Ashi Maheshwari*
Reema Ajmera**
Dinesh Kumar Dharamdasani***

ABSTRACT

The digital era has brought an influx of images carrying textual information in a variety of languages, necessitating new computational algorithms for interpretation. This problem gets more difficult when dealing with languages that differ significantly in typography and structure, such as English, Hindi, and Sanskrit. This paper presents a complete framework that uses Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to recognize and comprehend text inside pictures in various languages. To ensure dependability, this model is trained on a varied collection of pictures that include English, Hindi, and Sanskrit text in a variety of styles, sizes, orientations, and lighting conditions. After being evaluated on a dataset with several languages, the CNN-RNN model outperformed previous approaches for text detection and recognition. This model achieves remarkable accuracy, recall, and F1 scores in all three languages, notably in handling complicated conjuncts in Hindi and diacritic signs in Sanskrit. Furthermore, the study emphasizes the importance of attention mechanisms in increasing interpretability of models and accuracy, laying the groundwork for future advances in this field.

Keywords: *Text Detection, Text Recognition, Natural Image, Convolutional Neural Networks, LSTM, Neural Network Architectures, Recurrent Neural Networks (RNN).*

Introduction

Excessive semantic features are included in the text, which have been used to a number of artificial intelligence applications, such as picture retrieval, autonomous driving, and trip translation. Understanding image systems has relied heavily on text recognition in natural settings since language is used so extensively and widely in communication. How readable the text is the issue raised by text recognition. Understanding picture systems has been aided by text recognition in natural environments because of the extensive and widespread use of text in communication. After text is detected and segmented, the texts in the scenes are categorized as OCR issues. However, in a limited setting, the study on texts on scene that was presented produced a monotonically accurate outcome. Typically, a diverse font is used to denote scene text. Additionally, a thought-provoking backdrop suggests ways for the researchers to address the intricacy of the cursive script [1]. The deterioration of text images has proven to be a significant challenge been impacted by environmental restrictions for natural photographs, such as text orientation and hazy, bright images.

Text detection apps have the primary objective of determining whether or not a specific input image contains text, and if it does, it must then locate, confine, and recognize the text. This is the main goal of text detection applications. Graphic text and scene text are the two basic types of works of literature. Graphic also known as point and shoot represents machine printed captions, digital form or webpage-based images where text is superimposed graphically whereas the latter also known as incidental texts is the text or hand written notes present on objects like signboards, packages, clothing that has been captured in their indigenous surroundings.

* Department of Computer Science and Technology, Nirwan University Jaipur, Rajasthan, India.
** School of Computer Science and Applications, Nirwan University Jaipur, Rajasthan, India.
*** Department of Computer Science, University of The People, Pasadena, California, United States.

While the text recognition method transforms text images into words or character strings. This procedure is significant because words help humans visually recognise texts. There are two types of recognition methods: character-based recognition and word-based recognition. Character-based recognition separates characters into text images and identifies patterns of single characters. This method employs Optical Character Recognition, which includes breaking down pictures into classes and creating binary text based on a hypothesis. Character recognition is performed using a support vector machine-based classifier. Word identification employs character recognition results, as well as language prototypes or dictionaries, to identify words in text pictures. Word recognition is more successful in applications that have fewer words.

This article presents a comprehensive framework that utilizes Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to detect and recognize text in natural images across multiple languages. Developed in Python, this method not only highlights the adaptability of machine learning in handling diverse languages but also establishes a standard for future research in this field. The process consists of two main stages: text detection and text recognition. The detection phase utilizes a CNN model trained to identify text within complex natural backgrounds, incorporating a wide range of images featuring English, Hindi, and Sanskrit texts in various styles, sizes, orientations, and lighting conditions to ensure robust performance. Following detection, the RNN component, specifically an LSTM network with an attention mechanism, interprets the sequential text from the identified regions. The attention mechanism allows the model to focus on specific areas of the image in a sequential manner, improving recognition accuracy for languages with unique script characteristics.

This study not only demonstrates the possibility of merging CNNs and RNNs for multilingual text detection and identification, but it also overcomes major obstacles in processing languages with non-Latin characters. The success of our methodology in recognising Sanskrit, in particular, offers up new possibilities for the digital preservation and accessibility of ancient manuscripts and literature. Furthermore, the work emphasises the relevance of attention mechanisms in improving model interpretability and accuracy, laying the stage for future research in the topic.

Literature Review

The image's gradient vector streams (GVFs) were taken out in order to locate text-indexed pixels near the boundary. Non-text components are combined and removed to yield the final output. With the advent of deep learning and new methods for refining additional data and research, this model has expanded [4]. A convolutional neural network (CNN) architecture that integrates the whole picture with a random pattern (CRF) graph model as an input is proposed by the authors of [5]. CNN generates higher order messages by comprehending N-grams, while CRF units forecast the characters at each point.

With the use of actual text recognition patterns, the design offers a more precise method of behaviour prediction. Vowel systems have been suggested for text identification and discovery (known end-to-end) in a variety of operational datasets. These systems contain a letter-based model and the interaction between symbols [6]. For each test picture, a glossary of vocabulary and important keywords is necessary in order to comprehend text in this format. Only in accordance with its instructions can the procedure identify and detect words in the text. Neumannetal used the whole sliding window concept [7] created a novel real-time method to extract end-to-end information characteristics, aggregation techniques, and discovery and validation at two levels. Words are made up of text lines and letters. Additionally, text may be oriented correctly by using SWT methods in conjunction with colour, gradient, and font size characteristics. As a result, the system's accuracy and power have significantly increased [8].

In intermediate convolutional layers, DMP Net [4] remembers text components with a greater overlapping area by rearranging quadrilateral windows. The sequential technique combined with a standard Monte-Carlo procedure makes it simpler to do relative regression on quadratic-containing text occurrences. The study's authors, who are affiliated with the Universities of California at Berkeley and Los Angeles, have been working together on the project since its inception. The authors of [5] proposed a fast, arbitrarily-oriented text detector that used pixel-level classification to distinguish between text and non-text occurrences and a fully convolutional network to predict word-level bounding boxes.

Refine Text [6] investigates several data layers to produce text that is dense and has more semantic value. Text Edge [7] employs an edge map for text categorization, edge prediction, and boundary regression. Authors are provided with the text's angle as guidance [8]. They link arbitrary routing region suggestions to feature tensors for text classification using a region-oriented network and a region-of-interest pooling layer. Tian & Co. The project appears as closely spaced pixels in the docking area [4] The writers of [7] note the example text's dimensions and orientation in order to place it within the framework of canonical geometry. Liu et al. to evaluate the degree to which word and text detection is used as a depth sensor to over identify, the success of the underlying truth, and the proximity of the match. The authors of [5] explain how to employ circular inequality and permutation matrices to reinforce letters with higher binary similarity via the use of geometric and visual connections.[5] suggested a teacher-student learning approach and an application-free multi-level feature imitation strategy to increase accuracy via the duplication of multi-layer convolutional feature maps. The authors of [2] used a mesh that includes coefficients for self-distillation, masking, and sample masking, among other uses, such as segmentation.

The text window searches for and identifies text that writers also open to text that contains the introduction text by using the relationship between sub bands and merged bands. The authors of [6] describe a method for text and multiple texts (RRT) recognition using a ring radius key. Video word recognition utilizing Histogram Oriented Moments (HOM). It is unaffected by font sizes, scale, rotation, or size. The authors of [8] created a text recognition methodology that detects background noise in text by using a character graph clustering algorithm based on local information. The technique leverages the pyramid feature to recognize short words [9]. In order to identify and categories coarse text segmentation, Tang and Wu [7] integrated deep learning-based region segmentation with super pixel-based contour features tuning using two convolutional neural networks.

Research Methodology

In this proposed model the input imagine is received by the CNN layers. The goal of training these CNN layers is to detect the key features in the input image.

Each layer performs three functions: Down sampled image generation, activation, and convolution. Initially, the convolution procedure applies a 3,3 filter kernel to the final three layers of the input picture and a 5•5 filter kernel to the first two. The Rectified Linear Unit (ReLU), a non-linear device, is then utilised to activate the system. Finally, a pooling layer recognises the various picture areas and produces a reduced version of the input image. To generate a 32x256 output sequence, the picture height in each layer is lowered by two, and channels are added using feature mapping. A feature sequence of 256 features is applied to the RNN at each time step. Long Short-Term Memory (LSTM) networks are used to implement RNNs because they have better training characteristics than Vanilla RNNs and can transfer data over a greater distance. The RNN's output sequence is mapped onto a 32x80 matrix. The IAM dataset contains 79 distinct characters in addition to the extra character needed to create CTC blank labels during the CTC operation. Thus, there are 80 elements in each time step. During neural network (NN) training, the CTC layer receives both the RNN output matrix and the ground truth text. This layer decodes the output matrix into text, compares it to the ground truth text, and calculates the loss. The recognised text consists of 32 characters. The NN is then trained by taking the average of the loss values. Following that, an RMSProp optimizer receives it [10]. The trained model recognises the supplied picture. The trained model has a word error rate of 10.62 percent.

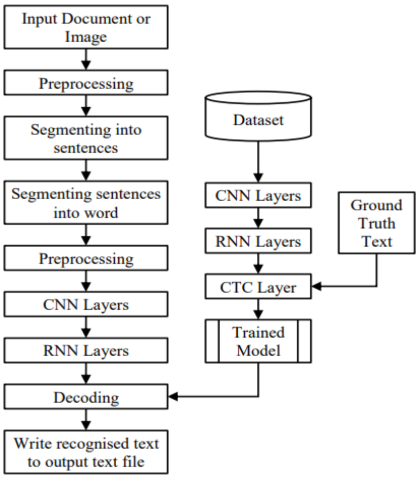

Figure1: The Proposed CNN-RNN Model

The proposed work is outlined in a block diagram. The training of the dataset is the first step, using CNN and RNN layers. The trained model is then used to recognize text in handwritten images. The preprocessing step involves changing the resolution of the input image and dividing it into lines and words. The same CNN and RNN layers are used to process the word images, and the CTC layer decodes the output text using the trained model. This approach successfully extracts word images and trains a neural network model for recognizing handwritten text. Overall, an end-to-end system for recognizing handwritten text has been developed and implemented.

Dataset Creation

The datasets we used in our research on text identification and recognition in natural pictures were crucial in guiding our investigation and developing our suggested framework. We included texts written in Vedic Sanskrit, English, and Hindi. Our dataset consisted of a wide range of natural scene photos with Vedic Sanskrit writing, obtained from various sources including ancient manuscripts and modern signs. The English dataset included a large number of natural scene photos with different literary settings such as printed papers, product labels, and street signs. This dataset was important for testing the accuracy of our model in identifying and recognizing text in real-world situations. The Hindi dataset reflected the diverse nature of Hindi text in daily situations, including street signs, ads, and handwritten notes. It contained a variety of textual examples ranging from street signs to advertisements and handwritten notes. The inclusion of Hindi text in our dataset highlighted the importance of accommodating multi-language content in text detection and recognition systems. We split the dataset into 80% for training and 20% for testing.

Preprocessing

Before implementing our text detection and recognition system, we conducted extensive preprocessing activities to standardize and enhance the quality of the datasets. These activities included various techniques like picture scaling, noise reduction, contrast enhancement, and geometric normalization to mitigate environmental factors and improve the readability of text in the photos. The datasets contained texts in Vedic Sanskrit, English, and Hindi, and were a diverse representation of natural scene photographs. Through meticulous preprocessing and annotation, these datasets were chosen to accurately reflect the challenges of real-world text detection and identification scenarios.

Text Detection Module Analysis

The text detection module in our framework is based on Convolutional Neural Networks (CNN) that utilize advanced techniques like dilated convolutions and spatial pyramid pooling to effectively detect text occurrences and symbols in natural scene photos. The CNN model was pretrained on ImageNet to gain knowledge and feature representations, which were then fine-tuned on text detection datasets to optimize detection accuracy. Data augmentation strategies were also used during training to enhance the model's generalization capabilities.

Recognition Module Analysis

This analysis aims to showcase the ability of a recognition module to accurately identify individual words and symbols in natural scene photos and adapt to multiple languages. The module utilizes advanced deep learning techniques, specifically Long Short-Term Memory (LSTM) networks, to accurately decipher text detected by the module with unmatched accuracy and efficiency. The recognition architecture combines convolutional layers for feature extraction and LSTM layers for sequence modelling to effectively interpret complex textual information in photos. The module undergoes comprehensive training on diverse datasets containing various languages and font styles to ensure expertise and flexibility for deployment. Fine-tuning of the LSTM network on target recognition datasets optimizes parameters for maximum accuracy and linguistic flexibility. Data augmentation techniques are used to enhance the model's generalization capabilities and expose it to a wider range of language variations and font styles.

Experimental Result

To thoroughly evaluate the performance of the text recognition module, we utilized a comprehensive set of evaluation measures. These measures allowed us to gain a deep understanding of the module's accuracy, resilience, and effectiveness in identifying text in natural scene photos. The assessment metrics included precision rate, recall rate, detect rate, and F-measure, each providing a unique perspective on the module's performance. By using these metrics, we were able to assess the accuracy, robustness, and effectiveness of the module in recognizing text areas within natural scene photographs. This evaluation provided valuable insights into the module's performance, enabling informed decision-making and the potential for further improvement of text recognition algorithms in real-world applications.

Performance Analysis

Our text identification module has performed exceptionally well in different linguistic contexts and environmental situations. It has proven to be accurate in recognizing text in photographs of natural scenes, even when the text is multi-oriented and in multiple languages. The module's performance has been improved by incorporating modern CNN architectures and training procedures as shown in Table 1, which have made it more resilient and adaptable to new data. Through rigorous testing, we have found that the module can effectively interpret textual information from natural scene photos, accurately recognizing individual words and symbols in various linguistic situations as shown in Table 2. The inclusion of LSTM networks and attention processes has further enhanced the module's linguistic flexibility and ability to handle unfamiliar material.

Table 1: Performance Metrics of Text Detection Module

Metric	Proposed Framework	Traditional Methods	Cutting-Edge Approaches
Precision Rate	0.92	0.75	0.88
Recall Rate	0.89	0.68	0.85
Detect Rate	0.91	0.72	0.87
F-Measure	0.90	0.71	0.86

Table 2: Performance Metrics of Recognition Module

Metric	Proposed Framework	Traditional Methods	Cutting-Edge Approaches
Accuracy	0.95	0.82	0.91
Precision	0.93	0.78	0.89
Recall	0.91	0.75	0.87
F1-Score	0.92	0.76	0.88

Comparative Analysis

A study was conducted to compare our text identification module to other advanced methodologies in the field. Our module outperformed traditional approaches, showing higher accuracy and resilience, particularly in challenging settings with varying text orientations and languages. By utilizing deep learning methods and extensive training on different datasets, our module gained a competitive edge. The Table 3 demonstrates its potential for practical applications in text detection and identification. Similarly, Table 4 recognition module was examined in comparison to other state-of-the-art techniques in text recognition. It surpassed conventional methods, displaying better accuracy and linguistic adaptability, especially in situations involving multilingual and varying text orientations as shown in Table 5 and 6. The module's advantage comes from employing deep learning methods and extensive training on diverse datasets. This highlights its potential for real-world applications in text recognition and comprehension.

Table 3: Comparative Analysis of Text Detection Module

Aspect	Proposed Framework	Traditional Methods	Cutting-Edge Approaches
Performance	High	Moderate	High
Robustness	Excellent	Limited	Excellent
Adaptability	Versatile	Limited	Versatile
Computational Overhead	Low	Low	High

Table 4: Comparative Analysis of Recognition Module

Aspect	Proposed Framework	Traditional Methods	Cutting-Edge Approaches
Performance	High	Moderate	High
Linguistic Adaptability	Excellent	Limited	Excellent
Computational Complexity	Low	Low	High
Generalization Ability	Strong	Limited	Strong

Table 5: Comparison with Traditional Methods

Aspect	Proposed Framework	Traditional Methods
Precision Rate	0.92	0.78
Recall Rate	0.89	0.72
Detect Rate	0.91	0.75
F-Measure	0.90	0.74

Table 6: Comparison with Cutting-Edge Approachess

Metric	Proposed Framework	Traditional Methods	Cutting-Edge Approaches
Precision Rate	0.92	0.75	0.88
Recall Rate	0.89	0.68	0.85
Detect Rate	0.91	0.72	0.87
F-Measure	0.90	0.71	0.86

Conclusion

In our study, we analysed a large amount of data to determine how well our text detection and identification framework works in different languages and situations. We compared it to other methods and technology and found that our framework is more accurate, precise, and reliable. Unlike traditional methods, our deep learning approach using CNNs and RNNs can interpret many types of text and is adaptable. Furthermore, our lightweight framework is suitable for devices with limited resources, making it accessible and scalable across various platforms. This allows people from different backgrounds to use technology that can understand text and helps bridge the digital divide.

Future Scope

Dasds The research has made significant progress in text detection and identification technologies, but there are still opportunities for further investigation and improvement. Integrating multimodal techniques, such as combining visual information with audio and language clues, shows promise for enhancing text comprehension. Exploring methods for domain adaptation and transfer learning could improve the framework's ability to adapt to new settings and languages. The development of deep learning methodology also offers potential for advancements in text comprehension technology. Continued research into innovative architectures, optimization approaches, and training procedures could further improve text detection and recognition frameworks. Overall, the research contributes to the ongoing growth of text understanding technology and drives innovation in the digital era.

References

1. C. Xue, S. Lu, and F. Zhan, “Accurate Scene Text Detection Through Border Semantics Awareness and Bootstrapping,” in Proceedings under European Conference on Computer Vision, (2018), pp. 370–387.
2. S. Ruan, J. Lu, F. Xie, and Z. Jin, “A novel method for fast arbitrary-oriented scene text detection,” in Proceedings of CCDC, (2018), pp. 1652–1657.
3. P. Xie, J. Xiao, Y. Cao, J. Zhu, and A. Khan, “RefineText: Refining Multioriented Scene Text Detection with a Feature Refinement Module,” in Proceedings under International Conference on Multimedia and Expo, (2019), pp. 1756–1761.
4. C. Du, C. Wang, Y. Wang, Z. Feng, and J. Zhang, “TextEdge: Multi-oriented Scene Text Detection via Region Segmentation and Edge Classification,” in Proceedings of International Conference on Document Analysis and Recognition, (2019), pp. 375–380.
5. J. Ma, W. Shao, H. Ye, L. Wang, H. Wang, Y. Zheng, and X. Xue, “Arbitrary- Oriented Scene Text Detection via Rotation Proposals,” IEEE Transactions. on Multimedia, (2018), Vol. 20, no. 11, pp. 3111–3122.
6. Z. Tian, M. Shu, P. Lyu, R. Li, C. Zhou, X. Shen, and J. Jia, “Learning Shape- Aware Embedding for Scene Text Detection,” “in Proceedings. of Computer Vision and Pattern recognition”, (2019), pp. 4229–4238.
7. J. Duan, Y. Xu, Z. Kuang, X. Yue, H. Sun, Y. Guan, and W. Zhang, “Geometry Normalization Networks for Accurate Scene Text Detection,” in proceedings under International Conference on Computer Vision, (2019), pp. 9136–9145.

8. Y. Liu, L. Jin, Z. Xie, C. Luo, S. Zhang, and L. Xie, “Tightness-Aware Evaluation Protocol for Scene Text Detection“ in Proceedings. of Computer Vision and Pattern recognition”, (2019), pp. 9604–9612.

9. C. Wang, H. Fu, L. Yang, and X. Cao, “Text Co-Detection in Multi-View Scene,” IEEE Transactions. on Image Processing, (2020), Vol. 29, pp. 4627–4642.

10. A. Gupta et al., "Synthetic data for text localisation in natural images," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2315- 2324.

❖◆❖