A Study on a Hybrid CNN-RNN Model for Handwritten Recognition Based on Deep Learning

Jun Ma
College of Economic and Management, Beijing Jiaotong University, Beijing, China

Keywords: Offline Handwritten Recognition, Hybrid CNN-RNN Network, Convolution Neural Network, Deep Learning.

Abstract: In today's era of digitalization, the efficient conversion of handwritten content into digital formats remains essential despite the widespread adoption of digital document storage. This study addresses the pressing need for efficient conversion of handwritten content into digital formats. Furthermore, the preprocessing procedures employed on handwritten images, including deskewing and normalization, were delineated. This study embraces a hybrid model-oriented recognition approach by utilizing the proposed hybrid Convolutional Neural Network (CNN)-Recurrent Neural Network (RNN) model for handwritten text recognition. It juxtaposes it with a solitary CNN model. The hybrid model's central components include a CNN for feature extraction and a Bidirectional Long Short-Term Memory network for sequence modeling. These components work together to enhance the precision of recognizing handwriting text. The research employs visualization techniques to understand the model's operations and improve performance. The CNN-RNN hybrid model significantly outperforms the CNN model, achieving a 12.04% reduction in Word Error Rate (WER) and a 5.13% Character Error Rate (CER). Conclusions drawn from the study illustrate that the suggested hybrid deep neural network model outperforms the conventional CNN method in terms of handwriting recognition accuracy. This is conducive to advancing the practical application of handwritten text scanning and recognition.

1 INTRODUCTION

In the contemporary age of digitalization, the adoption of digital document storage has become prevalent, encompassing various domains ranging from educational institutions to corporate settings. Nevertheless, despite the rapid progress in digital technology, which has made handwritten input on electronic platforms seamless and user-friendly, a substantial volume of paper-based documents, including manuscripts, contracts, invoices, and more, retains its significance. Consequently, there is an urgent necessity for efficient and precise conversion of handwritten content into digital formats to facilitate organized storage, effective management, and effortless retrieval. Handwriting recognition technology is increasingly acknowledged as a highly effective solution for enhancing work efficiency and systematically digitizing paper documents, thanks to its accuracy and versatile applicability. The handwriting Recognition research has two basic ways to explore: Online Recognition and Offline Recognition. The online recognition system actively monitors the generation of handwritten strokes in real-time when users utilize digital pens and electronic screen devices for writing, accurately identifying characters or words. This recognition process is typically performed pen-by-word or word-by-word, dynamically occurring as the user writes. In contrast to the real-time nature of Online Recognition, Offline Handwriting Recognition refers to a post-completion recognition system specifically designed to analyze handwritten content after it has been entirely generated. It is employed for recognizing scanned computer images containing handwritten documents, utilizing image processing techniques and pattern recognition algorithms to effectively analyze and identify handwritten characters, words, or sentences. Handwritten character recognition has many applications, from document recognition in digital offices to handwritten prescription storage in the medical field. It plays a crucial role in preserving cultural heritage by enabling handwritten font recognition in the digital archiving of historical documents, handwritten letters, ancient manuscripts, and other valuable cultural artifacts.

Ma, J.
A Study on a Hybrid CNN-RNN Model for Handwritten Recognition Based on Deep Learning.
DOI: 10.5220/0012801000003885

In *Proceedings of the 1st International Conference on Data Analysis and Machine Learning (DAML 2023)*, pages 268-273
ISBN: 978-989-758-705-4

As research in artificial neural networks has deepened, many deep learning methods, including Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) have become enriched and matured and have gained significant attention as promising research directions in the field of handwritten font recognition. Specifically, Artificial Neural Network (ANN) has an advantage in pattern recognition for handwritten characters (Abiodun et al 2019), and it is competent for ANN to handle non-linear tasks. Regarding CNN (Vaidya et al 2018), it can successfully capture the tiniest information in images, facilitating feature extraction for handwritten characters. CNN models leverage parameter sharing, reducing the number of learnable parameters and streamlining deep network training. Moreover, their ability to capture spatial hierarchies of features enables CNN models to excel in identifying stroke patterns, connectivity, and overall font structural characteristics. In contrast, using RNNs, especially LSTM units as fundamental structures further enhances recognition capabilities by considering sequential information and dependencies within the handwritten text. Adapting an LSTM structure allows RNNs to effectively maintain and update internal states when dealing with long sequences in time-series machine-learning tasks (Abiodun et al 2019). This ability to capture temporal dependencies is particularly beneficial in deciphering the nuances of handwritten fonts. The advantages of deep learning, such as scalability, adaptability to diverse data, and improved recognition accuracy, make them pivotal in the ongoing progress of handwritten font recognition research. This paper endeavors to enhance the model's performance through an initial deskewing of handwritten images, followed by the application of various techniques such as cropping and scaling. These techniques effectively adapt the images to a size and style conducive to utilization as input to the model. Additionally, to enhance the model's precision in identifying handwritten fonts, I adopted a CNN-RNN hybrid model to extract image features and complete the classification problem of word recognition. In addition, a single CNN method is used for comparison to highlight the improvement in recognition of the CNN-RNN hybrid model.

The paper is structured in the following manner: Section II highlights the contribution of related works in the handwriting recognition field. Section III demonstrates the fundamental architecture and principles of the CNN-RNN hybrid model. Section IV describes the entire experimental designation and the outcomes of model performance. Section V pertains to the overall conclusion and suggests feasible actions for future research.

2 RELATED WORKS

Significant advancements in handwriting recognition have been made through the emergence of innovative approaches in recent research. Geetha et al. introduced a hybrid model that employs a CNN that can grasp tiny featural image details and an RNN-LSTM to improve recognition precision (Geetha et al 2021). More recently, an effective method has been proposed to generate similar image samples with arbitrary lengths from original handwriting recognition samples (Kang et al 2022). This alleviates the problem of manually annotating handwritten data and enables training the handwriting recognition model with smaller image samples. Additionally, Gupta and Bag introduced a polygonal approximation-based approach for Devanagari character recognition, validated using multiple neural networks (Gupta and Bag 2022).

Furthermore, Zouari et al. presented a fusion model that utilized beta-ellipse parameters of segmented handwritten characters, combining TDNN-SVM for clustering and training, yielding impressive outcomes on extensive multilingual datasets (Zouari et al 2018). Carbune et al. described a multilingual system that services online handwriting based on LSTM architecture in conjunction with Bézier curves (Carbune et al 2020). In another study, Alam et al. developed a hybrid recognizer that combines LSTM and CNN models for recognizing writing trajectories in motion gesture digit and letter datasets (Alam et al 2020). Lastly, Zhang et al. explored a tree-BLSTM architecture, a variant of the LSTM model, to recognize two-dimensional mathematical expressions (Zhang et al 2020).

3 METHODS

3.1 Handwriting Image Preprocessing

Inspired by Hu's seven-moment invariants, a method that uses all pixel information to calculate the center distance is adopted (Devi and Amitha 2014). The image skew is determined for each handwritten text image by dividing the second-order central moment by the first-order central moment with regard to pixels and intensities. The image's skew is then rectified through the inverse mapping technique alongside linear interpolation via an affine transformation, and the correction matrix is computed and included within the skew parameter. After the corrected image is

normalized using the Z-score for each channel, it is transformed into a suitable input format for the hybrid model.

3.2 CNN-RNN Hybrid Deep Learning Model

The core of the convolutional layer for feature extraction in CNN is to use a convolutional kernel to slide over an image, capturing subtle patterns in different regions. The specific configuration of convolutional kernels in each convolutional layer introduces translation invariance, allowing consistent features to be extracted from various positions within handwritten characters.

Deep CNNs are typically used to extract and capture abstract representations and high-level features such as handwritten strokes (Vaidya et al 2018). However, as the depth increases, the gradient vanishing problem becomes the biggest obstacle. The residual network effectively solves this problem by introducing residual blocks and using skip connections across specified layers. This innovative architecture allows gradients to flow more freely during training, facilitating the training of very deep networks without suffering from vanishing gradients.

This study employs a deep neural network architecture rooted in ResNet-34 (He et al 2016). The ResNet-34 is a CNN model that is well-known for its residual architecture. It was trained through the ImageNet dataset and has gained widespread recognition in the field of computer vision. This weight-retention model serves as the foundational architecture for extracting features within the CNN framework in this study. Employing transfer learning to retain existing weights also facilitates rapid and efficient training on limited-scale datasets. I modified the ResNet-34 architecture by removing its classification output layer and incorporating an adaptive average pooling layer. I subsequently shaped the pooled output into a tensor form 96*1 by applying pooling to the feature tensor that ResNet-34 processed. This manipulation aims to create a more suitable feature representation for the subsequent RNN model, which is utilized for time series recognition.

3.3 Sequence Modeling with Bidirectional LSTM

RNNs leverage their short-term memory processing capabilities within their internal state to effectively handle the sequential nature of data. This property makes RNNs particularly valuable for addressing segmentation tasks involving time series data, such as handwriting recognition (Geetha et al 2021).

This study adopts the LSTM cyclic neural network structure, and the internal state is reasonably updated by introducing a gating mechanism to capture long sequence information effectively. An LSTM can be seamlessly integrated with CNN in an end-to-end fashion through the reshaped operation toward tensors. By configuring a stack of four hidden LSTM layers, the LSTM efficiently captures long-term relationships between successive strokes by working on the CNN's output tensor along the time dimension. Additionally, LSTM in bidirectional mode captures both forward and backward information flow in time, enhancing contextual understanding within the sequence. Next, the bidirectional LSTM output is implemented by a 1*1 convolution kernel to perform operations similar to linear transformations. It maps the hidden state at each time step to a vector of output size features depending on statement length without changing the relationship between time steps.

4 RESULTS

This study conducted a selected handwriting data set to test the proposed CNN-RNN fusion model. The computers used in this case study have the following configurations: i5-12500H 2.50 GHz CPU; 16 GB RAM; RTX-2050 4 GB GPU.

This study employed the IAM Handwriting Database 3.0, which contains handwritten English text contributed by 657 authors. The dataset comprises 5685 complete sentences, a total of 115,320 words, and 13353 lines. The XML file provided by this dataset uniquely corresponds to the text content in the captured image fragments, serving as labels. In the IAM dataset, which includes 26 English letters, 27 punctuation marks, and spaces, all characters fall within the recognition range of this handwritten font. It is worth noting that all letters in the recognition results of the IAM dataset are presented in lowercase. Figure 1 shows scanned images of some handwritten sentences in the IAM dataset using different writing styles and shows the textual content of the handwritten words in each image.

Two common indicators were employed to assess the model's performance, which includes Word Error Rate (WER) and Character Error Rate (CER):

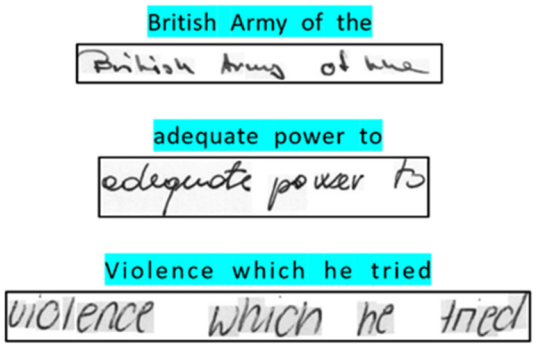

Figure 1: Scanned handwriting images and their corresponding text content in the IAM dataset (Picture credit: Original).

In (1), the equation is a metric used to assess a model's accuracy in recognizing complete words (Dutta et al 2018). It calculates the error rate by comparing the words generated by the model with those in the actual text, which is regarded as a crucial evaluation metric in multiple tasks. It is typically expressed as a percentage, with lower rates indicating better model performance.

$$WER = \frac{\sum_{i=1}^{N} EditDisance(GT_i, PT_i)}{\sum_{i=1}^{N} \#Words(GT_i)} \quad (1)$$

Equation (2) measures a model's accuracy in recognizing individual characters (Dutta et al 2018). It quantifies the extent of mismatching the predicted and actual words and characters. The editing distance for all samples is considered and divided by the number of existing characters, forming a ratio representing the extent of prediction error. The lower the edit distance, the lower the CER, and the higher the recognition accuracy of the representation model at the character level.

$$CER = \frac{\sum_{i=1}^{N} EditDisance(GT_i, PT_i)}{\sum_{i=1}^{N} \#Chars(GT_i)} \quad (2)$$

All data underwent two rounds of skewness data processing, and some of the processing results are visualized as shown in Fig. It shows the changes for selected words after deskewing. The top row is the handwritten word image scanned in the original data set. The middle row is the image obtained after one deskewing process, and the bottom row is the handwritten font image after two deskewing procedures. Deskewing can reduce recognition errors caused by skewing individual letters within the sentence to a certain extent. Before model training, the processed and normalized images were resized to 224*224 px, which facilitates data input for subsequent model training. Next, all data and the related word list were split; 80% of the handwritten images and their labels in the data set were divided into training sets for model training, while the remaining 20% were used as test sets to measure the prediction results of the model.

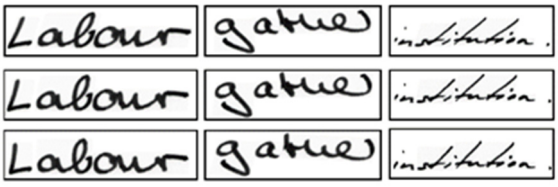

Figure 2: Original handwritten images and corresponding images after one and two deskewing processes (Picture credit: Original).

The hybrid CNN-RNN and comparative CNN models use the processed data for training and prediction comparison. The applied CNN model maintains the core architecture of the original ResNet34, aligning with the CNN component of the CNN-RNN fusion model, except for the model output layer. ResNet34's backbone, initially pre-trained on ImageNet, was leveraged in our research to enhance model performance and expedite training. This transfer learning approach, retaining the original weights, enabled the creation of a task-specific model, effectively utilizing ResNet34's learned features from ImageNet, thereby boosting training efficiency and enhancing model performance.

The CNN part of the model has a 3-channel input layer that accepts RGB images of 224*224 px size as model input. The initial component of the architecture employs a 7*7 convolutional layer tasked with extracting foundational image textures and edge features. Subsequently, a profound network of residual convolutions emerges, comprising four consecutive residual blocks. Within this architectural framework, each convolutional kernel maintains dimensions of 3x3.

Figure 3 illustrates the feature map output obtained as the image passes through each convolutional layer (or residual block) during the CNN feature extraction process. After passing through the first convolutional layer, the RGB image yields 64 output channels, which are input into the subsequent residual block within the ResNet-34 convolutional neural network. The following four residual blocks, namely R1, R2, R3, and R4, produce 64, 128, 256, and 512 output channels, respectively.

In a single-method CNN model, the image is passed through the residual neural network, subsequently flattened, and then sent to the fully connected layer, where the output size is adjusted to generate the classification model's output. In contrast,

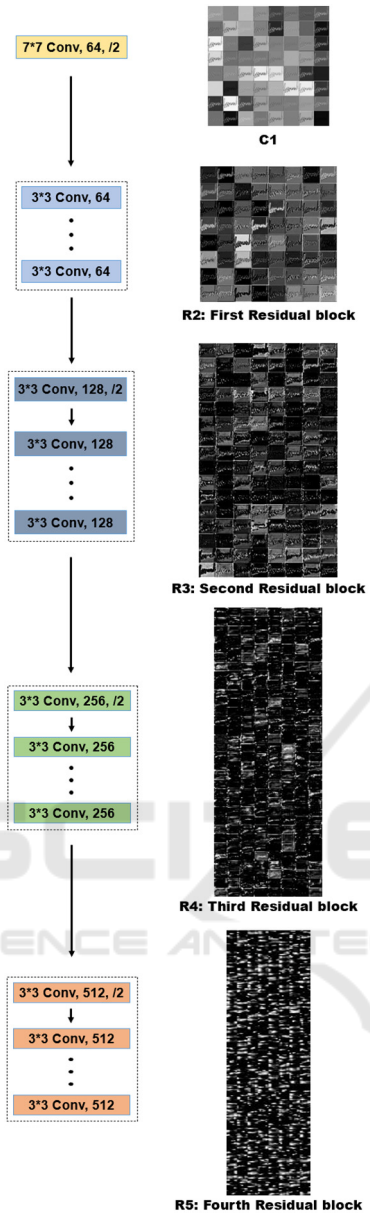

Figure 3: Feature map of residual block and convolutional layer output in residual convolutional neural network (Picture credit: Original).

for the CNN-RNN fusion model, the image is flattened and forwarded to the LSTM unit for further training.

Figure 4 demonstrates that the CNN-RNN fusion model reaches an overall WER of 12.04%, approximately 26% lower than the CNN single method. The hybrid CNN-RNN model exhibits significant improvements in character recognition, with a CER improvement of over 17%, ultimately achieving a CER of 5.13% when compared to the CNN model.

Figure 5 displays the prediction results for the scanned image and original text content. Most words and punctuation marks are accurately recognized, with no omitted letters. However, some individual letters in certain words may be incorrectly recognized.

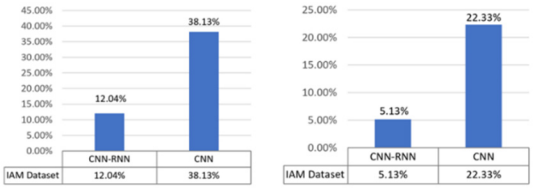

Figure 4: Comparison of recognition results between CNN and hybrid CNN-RNN models (Picture credit: Original).

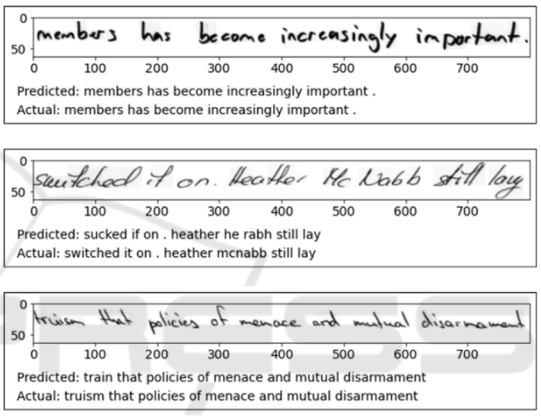

Figure 5: Comparison of handwriting recognition results and original content based on hybrid CNN-RNN model (Picture credit: Original).

5 CONCLUSION

This study proposes a solution for handwritten text recognition on the IAM dataset using a hybrid CNN-RNN model. A single-method CNN model will be employed as a comparative model to assess the impact of different models. The study utilized visualization techniques to demonstrate the various operations of the input image during each stage of the model. This aimed to enhance the understanding of the overall structure of the model as well as the processing details of its components. The results demonstrate a significant improvement in the CNN-RNN fusion model compared to the CNN model. The fusion model's optimal performance reduces WER to 12.04% and CER to 5.13%.

While this article proposes a superior fusion model for addressing handwriting recognition challenges, it still faces issues like insufficient recognition accuracy and a lack of case sensitivity in English characters. For

future research, consider exploring the CNN-RNN-CTC (Connectionist Temporal Classification) combined model, leveraging CNN to extract image feature details, RNN for sequence modeling, and CTC for sequence labeling to enhance recognition accuracy. Incorporating case sensitivity into the model can expand its applicability and improve recognition accuracy.

REFERENCES

O. I. Abiodun, M. U. Kiru, A. Jantan, et al. "Comprehensive review of artificial neural network applications to pattern recognition," IEEE Access, vol. 7, pp. 158820-158846, 2019.

R. Vaidya, D. Trivedi, S. Satra, and P. M. Pimpale, "Handwritten Character Recognition Using Deep-Learning," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018, pp. 772-775.

R. Geetha, T. Thilagam, and T. Padmavathy, "Effective offline handwritten text recognition model based on a sequence-to-sequence approach with CNN–RNN networks," in Neural Computing and Applications, vol. 33, no. 17, pp. 10923–10934, 2021.

L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Content and Style Aware Generation of Text-Line Images for Handwriting Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 8846-8860, 2022.

D. Gupta, and S. Bag, "Holistic versus segmentation-based recognition of handwritten Devanagari conjunct characters: a CNN-based experimental study," Neural Comput and Applic, vol. 34, no. 7, pp.5665-5681, 2022.

R. Zouari, H. Boubaker, and M. Kherallah, "Multi-language online handwriting recognition based on beta-elliptic model and hybrid TDNN-SVM classifier," Multimed Tools Appl, vol. 78, no. 9, pp. 12013-12123, 2018.

V. Carbune, P. Gonnet, T. Deselaers, et al. "Fast multi-language LSTM-based online handwriting recognition," International Journal on Document Analysis and Recognition (IJDAR), vol. 23, no. 2, pp. 89-102, 2020.

M. S. Alam, K. Kwon, M. A. Alam, M. Y. Abbass, S. M. Imtiaz, and N. Kim, "Trajectory-Based Air-Writing Recognition Using Deep Neural Network and Depth Sensor," Sensors, vol. 20, no. 2, pp. 376, 2020.

T. Zhang, H. Mouchère, and C. Viard-Gaudin, "A tree-BLSTM-based recognition system for online handwritten mathematical expressions," Neural Comput and Applic, vol. 32, no. 9, pp. 4689–4708, 2020.

S. S. Devi and T. Amitha, "Offline handwritten writer independent Tamil character recognition," International Conference on Information Communication and Embedded Systems (ICICES2014), 2014, pp. 1-6.

K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770-778.

K. Dutta, P. Krishnan, M. Mathew, and C. V. Jawahar, "Improving CNN-RNN Hybrid Networks for Handwriting Recognition," 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 2018, pp. 80-85.