



Article

Attention-Based CNN-RNN Arabic Text Recognition from Natural Scene Images

Hanan Butt ¹, Muhammad Raheel Raza ^{2,*} , Muhammad Javed Ramzan ¹ , Muhammad Junaid Ali ¹ 
and Muhammad Haris ¹

¹ Department of Computer Science, COMSATS University Islamabad, Islamabad 45550, Pakistan; sp18-rcs-023@student.comsats.edu.pk (H.B.); fa18-rcs-024@student.comsats.edu.pk (M.J.R.); sp18-rcs-028@student.comsats.edu.pk (M.J.A.); fa19-rcs-046@student.comsats.edu.pk (M.H.)

² Department of Software Engineering, College of Technology, Firat University, 23000 Elazig, Turkey

* Correspondence: 191137125@firat.edu.tr; Tel.: +90-5523958926

Abstract: According to statistics, there are 422 million speakers of the Arabic language. Islam is the second-largest religion in the world, and its followers constitute approximately 25% of the world's population. Since the Holy Quran is in Arabic, nearly all Muslims understand the Arabic language per some analytical information. Many countries have Arabic as their native and official language as well. In recent years, the number of internet users speaking the Arabic language has been increased, but there is very little work on it due to some complications. It is challenging to build a robust recognition system (RS) for cursive nature languages such as Arabic. These challenges become more complex if there are variations in text size, fonts, colors, orientation, lighting conditions, noise within a dataset, etc. To deal with them, deep learning models show noticeable results on data modeling and can handle large datasets. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can select good features and follow the sequential data learning technique. These two neural networks offer impressive results in many research areas such as text recognition, voice recognition, several tasks of Natural Language Processing (NLP), and others. This paper presents a CNN-RNN model with an attention mechanism for Arabic image text recognition. The model takes an input image and generates feature sequences through a CNN. These sequences are transferred to a bidirectional RNN to obtain feature sequences in order. The bidirectional RNN can miss some preprocessing of text segmentation. Therefore, a bidirectional RNN with an attention mechanism is used to generate output, enabling the model to select relevant information from the feature sequences. An attention mechanism implements end-to-end training through a standard backpropagation algorithm.

Keywords: image text recognition; deep learning; recurrent neural networks (RNNs); convolutional neural networks (CNNs); bidirectional RNN; attention mechanism; text segmentation; natural scene images



Citation: Butt, H.; Raza, M.R.; Ramzan, M.J.; Ali, M.J.; Haris, M. Attention-Based CNN-RNN Arabic Text Recognition from Natural Scene Images. *Forecasting* **2021**, *3*, 520–540. <https://doi.org/10.3390/forecast3030033>

Academic Editors: Walayat Hussain, Honghao Gao and Sonia Leva

Received: 12 May 2021

Accepted: 13 July 2021

Published: 20 July 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The research community has been working for many years to recognize printed and written texts [1]. Many applications to the text recovery system help blind people to navigate specific environments [2]. The text is filtered with textual contents on the images. The text has high-level scene information in natural pictures. The styles, font sizes, and text may be different. Anyone can view natural text scenes in advertising notes and banners. Text with a background of water, constructions, woods, rocks, etc., mostly has noise. The process of text recognition from natural images can be slowed down with these backgrounds [3]. This problem may be more prominent when lights are loud, blurred, and texts are directed in images of nature [3,4]. Here, this paper's primary focus is on the recognition of Arabic texts from natural scenes.

For most people, video is a crucial source of information—there are those who understand videos better than other sources of information. Many videos on social websites and TV channels are available. Storage technology can now save a large number of video data. Since the length and annotation of these videos are increasing, quick and productive recovery systems should be available to enable access to the information relevant to the videos. The most significant information in videos is the title. Anybody can understand the text written on a video. The recognition of video text offers help in analyzing the content of the video [5]. The text data on the videos are highly linked to the people in the video, and they are also linked to the events of the video.

Reading text from a printed document such as an essay differs to natural scene recognition [6]. This problem has brought attention to the area of natural scene recognition. Scene text recognition has become an essential module for many practical applications due to the advancement of mobile devices such as self-driving cars, smartphones, etc. [7]. A lot of work has been carried out in recognizing printed text in the form of Optical Character Recognition (OCR) systems. Still, they cannot deal effectively with natural scenes with factors such as image distortion, variable fonts, background noise, etc. [8]. Gur et al. [9] described the problems in the text recognition system. According to them, OCR tools are not sufficient for a complete solution; human expertise is also required. Arabic is the fourth most widely spoken language globally, but the work on word-level text recognition for the Arabic language is negligible [10]. The original image containing one word imagescan be seen in Figure 1.



Figure 1. Arabic news channel dataset.

Arabic text acknowledgment is an active field of research in pattern recognition today. Arabic printed documents have been a topic of study for many years in the research community. Character recognition [11] makes up most of the previous work on the Arabic script. Due to its cursive fashion and lack of annotated information, the Arabic script is challenging [12]. Additionally, the OCR script system is also challenging in regard to Arabic because most English OCR techniques are not very practical for Arabic. For Arabic scripts, many researchers have worked on word segmentation, but now, people work on free text segmentation recognition. Models such as the recurrent neural network are used for this sequence by sequence (RNN). The input to target labels in a series is transcribed in these models [13]. For handwritten English and Arabic text recognition, the RNN model is used in [14,15], respectively. For Arabic video text recognition, this paper uses the same sequence to sequence model. A great deal of work has been carried out in [16,17] to recognize video text. The availability of Arabic script datasets is minor compared to

English script datasets, but for video text recognition, two datasets are available: the ACTIV dataset [18] and the ALIF dataset [19]. Due to the lack of annotated data of Arabic script, the work on word-level text recognition is very scarce [20].

DCNNs are used for various tasks such as feature extraction, classification, and objection detection. However, the input and output size of this network is not fixed [21]. If the model receives input data in a series, a sequence of labels should be predicted. RNN models resolve the problem of inputs and outputs of variable sizes. To recall previous input, RNN introduces the concept of memory [22]; the information used by RNN-based models is usually converted into sequence–function vectors [23]. A DCNN is used to obtain the functional representations of the input image in the proposed study. These features are transferred to a bidirectional RNN. These characteristics are transmitted to the two-way RNN. The mechanism of attention is applied to the RNN. The mechanism for attention gives the results of each RNN output. Relevant information will arrive at the front when necessary, via the attention mechanism [24]. The results are predicted based on this attention model.

For natural scene images, several text recognition systems are proposed. Bissacco et al. [25] recognize characters for English scripts on a single basis, and then the DCNN model recognizes the characters detected. These models' limitation is that the exactness of the word's cutting and detection is not very accurate. Due to the complications, this is harder for Arabic script. As far as our knowledge is concerned, there is less work in Arabic text script recognition at the word level. Here, we use images of natural scenes; Using images of natural scenes involves three challenges, including environmental challenges scene complexity, noise, etc. The mages may be blurred, and the content's text may have variable fonts, guidelines, and text colors. The complex properties of an Arabic script, including its semi-cursive mode, letter shape, paws (Arabic part), dots, and similar letters in different situations present other challenges. Figure 2 is a representation of Arabic natural scene text.



Figure 2. Arabic natural scene text dataset.

Jain et al. [13] proposed a hybrid CNN-RNN model to recognize Arabic text in natural scenes and videos. In this model, an input image is converted into a feature map by the CNN layer. This layer splits the feature map column-wise (f_1, f_2, \dots, f_i) and transfers to the RNN layer. The RNN layer reads the feature sequences one by one (mean one time step at a time) and obtains a fixed dimension vector that contains all the information of input feature sequences. Then, a transcription layer uses this fixed vector to generate the output sequence. The main issues of this approach include the fact that all the knowledge of these feature sequences is stored in a fixed-length vector. During the backpropagation, the model has to visit all the information in a fixed vector. This will be more difficult if we have long sequences. Therefore, there should be a mechanism that deals dynamically with these

feature sequences (input sequences). The attention mechanism dynamically picks the most relevant information from the feature sequence and feeds it to the context vector; for this, it soft searches where the vital information is present in the input sequences [26]. During backpropagation, the model will only visit the critical information in the context vector. This paper presents an attention-based CNN-RNN model for Arabic image text recognition.

2. Literature Review

The Arabic language is extremely complex, due to its many properties. It is written cursorily from right to left; it has only 28 letters. Most of its letters, except six letters, are connected to the previous letter at the bottom of the line. One or more paws can be an Arabic word (part of the word). A letter can be linked to the letter or linked in the form of a paw to the letter. The letter form makes the script more difficult, and the shape of it depends on the location of the letter in the beginning, in the middle, or at the end [27]. Ligature fashion [28] is also used in Arabic scripts, which means that two letters connect and form a new letter that cannot be separated. These characteristics make the automatic recognition of Arabic scripts more difficult. A. Shahin presented a regression-based model to recognize Arabic printed text on paper [29], which could be linear or nonlinear. First, they thinned the text from the images and segmented the images into sub-words. Following that, a coding scheme was used to represent the relationship between word segments. Each form of the character and font type had a code in the coding scheme. Finally, the linear regression model validated the representations against a truth table. They tested the model on 14,000 different word samples, and the accuracy was 86 percent.

There are two recognition methods for document- and video-based OCR techniques: one is the free segmentation method, and the other is based on segmentation. Mostly, two approaches are used for recognition: the first one is the analytical approach, which segments a line or word into smaller units, and the second one is the global approach, which does not perform any segmentation; it takes the whole line or word as an input. Most of the Arabic text recognition methods are segmentation-based. For example, Ben Halima [30] uses an analytical approach for Arabic video text recognition. They perform binarization on Arabic text lines and then segmentation by projection profiles. They use the fuzzy-k-nearest-neighbors technique on handcrafted features to classify characters. Lwata [31] also employs the analytical approach for Arabic text recognition on various video frames. They use a space detection technique to segment the line into words, which are then segmented into characters. Finally, they use an adapted quadratic function classifier to perform character-level recognition. With these approaches, segmentation errors can grow indefinitely and contribute to recognition performance.

A. Bodour presented a deep learning model for Arabic text recognition in their work. Their work was mainly focused on the recognition of Islamic manuscripts of history. To enhance the quality of scanned images and for segmentation, several preprocessing techniques were used. After preprocessing, they used CNNs for recognition and achieved 88.20% validation accuracy. A deep learning model is also used in [9] for handwritten Arabic text recognition. Various preprocessing techniques were used for image noise reduction. They used a threshold method to segment the characters of words into distinct regions. Feature vectors were constructed by these distinct regions. They evaluated the model on a custom dataset gathered from different writers. The model achieved 83% accuracy.

In research works [32,33], a CNN model with three layers is presented for the recognition of handwritten characters of Arabic script. The datasets used for evaluation were AHCD [34] and AIA9k [35]. The model achieved 97.5% validation accuracy. Another deep learning model with MD-LSTM is presented in [36] for handwritten Arabic text recognition. The model used the CTC loss function. The purpose of that work was to evaluate the results of extending the dataset using data enhancing techniques and compare the performance of an extended model with other related models. The dataset used for training and evaluation is handwritten KHAT [37]. The model achieved 80.02% validation

accuracy. In [13], the authors presented a hybrid deep learning model for printed Urdu text detection from scanned documents. The model used a CNN for feature extraction and bi-LSTM for sequence labeling. The model used the CTC loss function. The APTI and URDU datasets [38] were used for evaluation. The model achieved 98.80% accuracy. To illustrate, Figure 3 is a flowchart to separate words from Arabic sentence images.

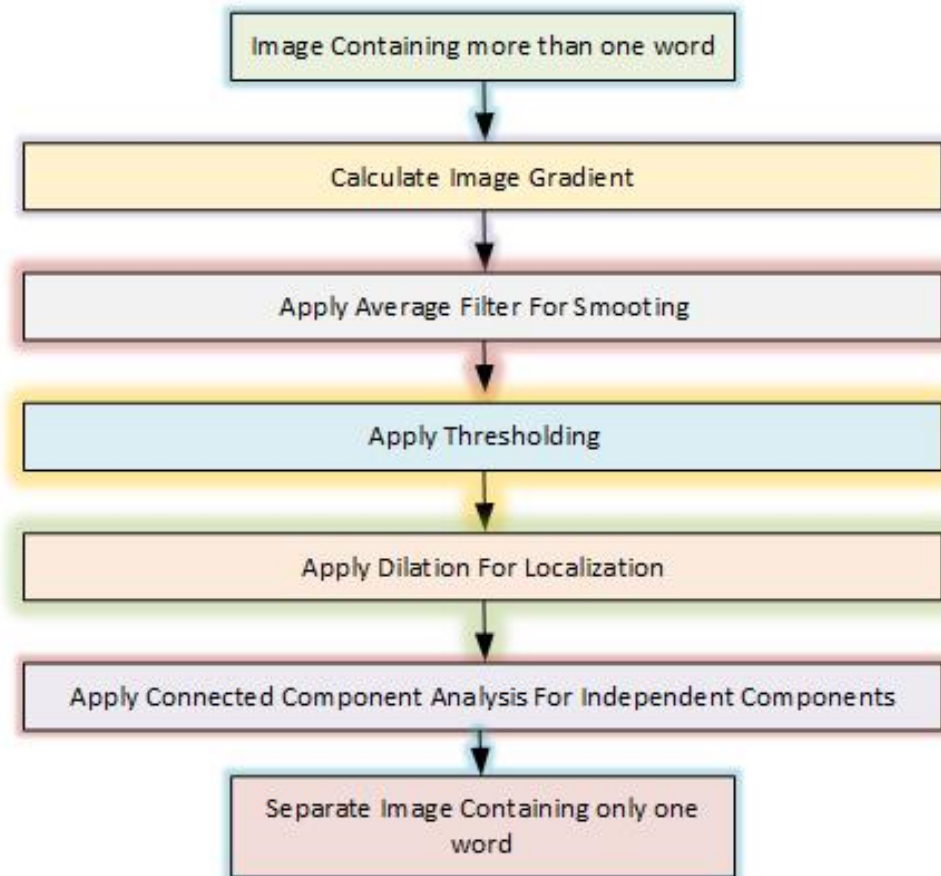


Figure 3. A flowchart of the proposed methodology used to separate words from Arabic sentence images.

Zhai [39] proposed a bidirectional RNN-based segmentation free model with the layer of CTC for the recognition of Chinese text, which was obtained from news channels. For model training, they composed almost two million news headings from 13 TV channels. There is a lot of work on automatic text recognition from natural scenes and videos for English scripts, but for Arabic scripts, the research community is newly active. For English scripts, Bissacco et al. [25] firstly detect the characters individually, then the DCNN model recognizes these detected characters. The limitation of these models is that the accuracy of cropping and detecting the character from the word is not high. This will be more difficult for the Arabic script due to its complications.

For English script text recognition, Jaderberg et al. [40] used a DCNN for detection but the recognition region mechanism was used. However, this model is not suitable for Arabic script due to its complications. The number of unique words in English script is less than in Arabic script. Almazán et al. [41] had two goals to fulfill: the first is word spotting and the second is word recognition on the images. The authors proposed an approach where text strings and word images both emerged to a common vector space. In semantic segmentation, RNN-based models are used to handle the long-distance pixel dependencies [42]. For example, in [43], Gatta et al. unrolled the CNN to obtain the semantic connections at different time steps. Authors [44–48] used LSTM with four blocks to scan all directions of an image; this helped to learn long range spatial dependencies.

There has been a lot of work performed in the field of text extraction from videos. Karray et al. [49] extracted text using classification methods. Hua et al. [50] used edge detection, which is a type of text extraction. Previously, the majority of work on Arabic videos was divided into two categories: dealing with relevant feature extraction and classifying features to target labels. To recognize Chinese handwritten text, authors [51–55] used a bidirectional LSTM network. The attention-based model inspired our proposed model. Papers [56–61] proposed different attention-based speech recognition model. Authors [62–71] proposed various attention-based models for object detection in images as well. Fasha et al. [72] proposed a hybrid deep learning model, and used five CNN layers and two LSTM layers for Arabic text recognition. They worked on 18 fonts of the Arabic language [73]. Graves et al. [74] created Neural Turing machines using an attention-based model with convolutions (NTMs). The NTM computes the attention weights based on content. Authors [75–77] present a supervised convolutional neural network (CNN) model that uses batch normalization and dropout regularization settings to extract optimal features from context. When compared to traditional deep learning models, this tries to prevent overfitting and improve generalization performance. In [78–80], for feature extraction, the first model employs deep convolutional neural networks (CNNs), whereas word categorization is handled by a fully connected multilayer perceptron neural network (MLP). SimpleHTR, the second model, extracts information from images using CNN and recurrent neural network (RNN) layers to develop the suggested deep CNN (DCNN) architecture. To compare the outcomes, Daniyal et al. presented the Bluechet and Puchserver models. Arafat et al. [81] introduce Hijja, a novel dataset of Arabic letters produced only by youngsters aged 7–12. A total of 591 individuals contributed 47,434 characters to our dataset. They propose a convolutional neural network-based automatic handwriting recognition model (CNN). Hijja and the Arabic Handwritten Character Dataset (AHCD) are used to train our model. Ali et al. present a method for recognizing cursive caption text that uses a combination of convolutional and recurrent neural networks that are trained in an end-to-end framework. The background is segmented using text lines retrieved from video frames, which are then fed into a convolutional neural network for feature extraction. To learn sequence-to-sequence mapping, the extracted feature sequences are fed to different forms of bidirectional recurrent neural networks along with the ground truth transcription. Finally, the final transcription is created using a connectionist temporal classification layer. Alrehali et al. present a method for identifying text in photographs of the Hadeethe, Tafseer, and Akhidah manuscripts and converting it into a legible text that can be copied and saved for use in future studies. Their algorithm's major steps are as follows: (1) picture enhancement (preprocessing); (2) segmenting the manuscript image into lines and characters; (3) creating an Arabic character dataset; (4) text recognition (classification). On three produced datasets, they employ a CNN in the classification stage. To solve the Arabic Named Entity Recognition (NER) task, El Bazi et al. developed a deep learning technique. They built a bidirectional LSTM and Conditional Random Fields (CRFs) neural network architecture and tested various commonly used hyperparameters to see how they affected the overall performance of our system. The model takes two types of word information as input: pre-trained word embeddings and character-based representations, and it does not require any task-specific expertise or feature engineering. Arafat et al. proposed a method for detecting, predicting, and recognizing Urdu ligatures in outdoor photographs. For identification and localization purposes, the unique Faster-RCNN algorithm was utilized in conjunction with well-known CNNs such as Squeezenet, Googlenet, Resnet18, and Resnet50 in the first phase for images with a resolution of 320 240 pixels. A unique Regression Residual Neural Network (RRNN) is trained and evaluated on datasets including randomly oriented ligatures to predict ligature orientation. A Two Stream Deep Neural Network was used to recognize ligatures (TSDNN).

According to a review of related work, less work exists for natural scene Arabic text recognition, particularly at the word level. Our performed work represents specific

steps in this research area and provides tools that can be used to expand and improve the results obtained.

3. Proposed CNN-RNN Attention Model

The architecture is divided into three sections: the first deals with convolutional layers, the second with recurrent layers of a bidirectional LSTM, and the third with recurrent layers of a bidirectional LSTM and an attention mechanism. The first part extracts feature representations from the input image. These representations are fed into the second part. This recurrent layer will label the sequences one by one. The connectionist temporal classification layer (CTC) is used when dealing with aligned data; it is essentially a loss function deal where time is a variable [48].

This paper uses two datasets, ACTIV and ALIF; they are firstly preprocessed, then fed to the convolutional layer. The size of all images is fixed before giving input to the first part of the architecture. Here, the height of each image is fixed. Convolutional layers will split the feature maps column-wise and create features sequence by sequence (f_1, f_2, \dots, f_i) . Due to CNN feature extraction, each feature contains important information. As CNNs cannot handle the long dependency, these features are directed as input to the second part of the architecture. This section makes predictions based on the features generated by the first section. In the second part, bidirectional Long Short-Term Memory (LSTM) networks are used. These networks deal with inputs and outputs in a recurrent fashion. The second part of the network will be unrolled based on a time variable, and the time variable will be determined solely by the input data. The disappearing gradient problem occurs in simple recurrent neural networks, but LSTM networks can learn meaningful information. They use different gates and memory cells to resolve the problem of vanishing gradient [49]. A bidirectional LSMT network is used in text recognition, as the learning pattern of text from right to left and left to right is very important [23]. The network can be made deeper by stacking these LSTM layers [50].

$$v = (v_1, v_2, v_3, \dots, v_t) \quad (1)$$

Now, 'v' has the information of both sides. The third part of the architecture has an LSTM network with attention, produces the results based on the information in 'v_i'. The attention mechanism gives scores to each feature sequence; for this, it gives probability to the recognized text.

$$Pr(y) = \prod_{t=1}^T Pr(y_t | \{y_1, y_2, y_3, \dots, y_{t-1}\}, con_t) \quad (2)$$

con_t is the context vector at time t generated from feature sequences. Probability with one LSTM layer is:

$$Pr(y_t | \{y_1, y_2, y_3, \dots, y_{t-1}\}, con_t) = ny_{t-1}, hs_t, con_t \quad (3)$$

Here, n is a non-linear function that gives the probability output of y_t . Here, hs_t is the hidden state of LSTM, which can be calculated as:

$$hs_t = func(hs_{t-1}, y_{t-1}, con_t) \quad (4)$$

The context vector can be calculated as:

$$con_t = \sum_t^{i=1} (W_{ti} * v_i) \quad (5)$$

Each feature sequence has a weight W_{ti} and can be calculated as:

$$W_{ti} = \frac{\exp(es_{ti})}{\sum_{i=1}^T \exp(es_{ti})} \quad (6)$$

es_i scores represent how well input around i related to the output at t . It can be calculated in the equation below. However, the proposed method is explained in Figure 4.

$$es_{ti} = \alpha(hs_{t-1} * v_i) \quad (7)$$

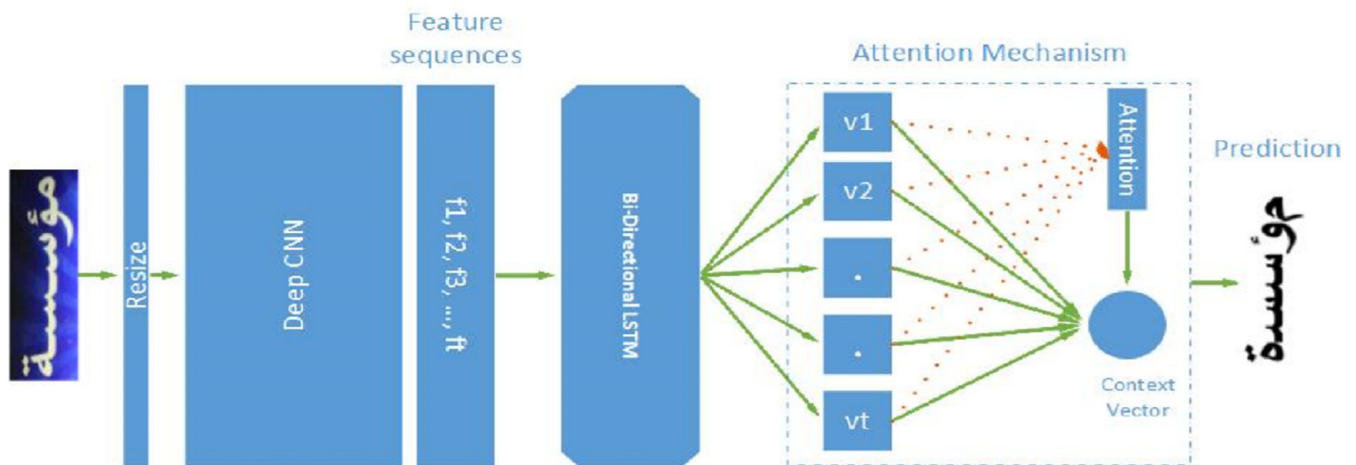


Figure 4. Proposed system model.

The proposed methodology uses an attention mechanism over a bidirectional LSTM; for this, it calculates the attention of each input sequence and then multiplies this attention to the respective input vector. The weighted sum of these attentions with their respective vectors is stored in the context vector. Attentions are the weights of input sequences.

3.1. Dataset and Preprocessing

The model was trained on two datasets, ACTIV and ALIF. There are 21,520 line images in the ACTIV dataset, and these images are from different news channels. The ALIF dataset is smaller than ACTIV, and has 6532 text lines images, and these are also from Arabic news channels. These images are annotated with XML files. Another dataset was curated by downloading freely available images containing Arabic script from Google Images and Facebook; this dataset contains cropped images of market banners, shop banners, etc. This Arabic dataset of 2477 images was manually annotated by experts. Tables 1 and 2 represent the statistical details of the ACTIV dataset and ALIF dataset. Details of the Google curated dataset is given in Table 3. Segmented word images can be seen in Figure 5.

Table 1. Statistical detail of ACTIV dataset.

TV-Channels	Training Data	Testing Data
	Lines Words	Lines Words
AL-JazeeraHD-TV	1725 7590	380 1286
Russia-Today-TV	2076 12,999	301 1946
France 24-TV	1821 5372	264 978
TunsiaNat-TV	1900 8950	290 1094
AIISD-TV	5840 26,203	812 5136

Table 2. Statistical detail of ALIF dataset.

TV-Channels	Training Data	Testing Data
	Lines Words	Lines Words
AL-JazeeraHD-TV	1430 5970	310 980
AL-Arabiya	410 2230	265 1457
France 24-TV	1690 4887	219 870
BBC-Arabic	155 350	95 180
AIISD-TV	5030 21,104	780 3013

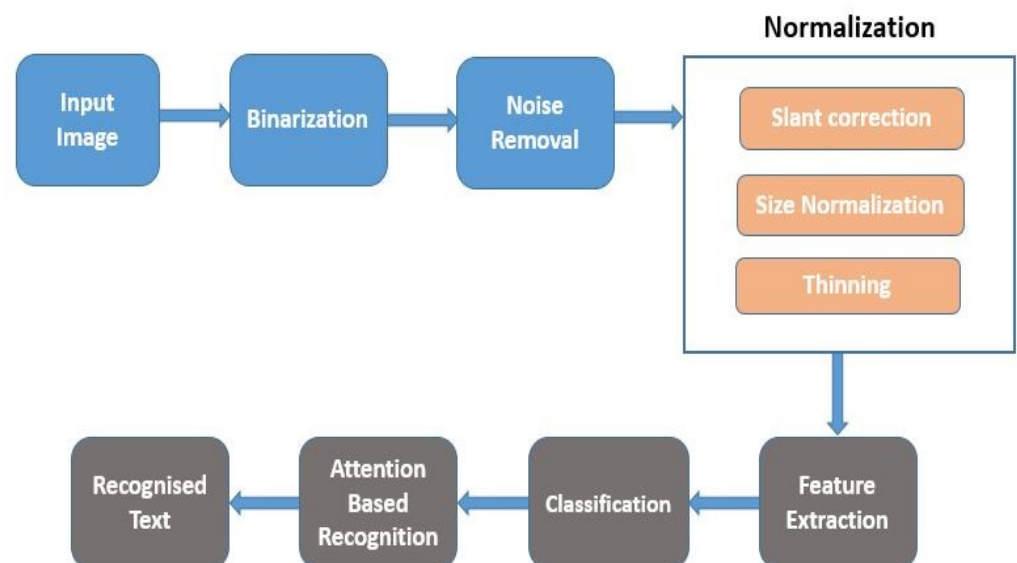
Table 3. Statistical detail of curated dataset.

Sources	Training Data	Testing Data
	Lines Words	Lines Words
Google	1100 3320	409 1110

**Figure 5.** Sample images from one word dataset.

3.2. Segmented Dataset

Here, we had two datasets, ACTIV and ALIF, and an image may have had more than one word. We worked on word-level text recognition; we needed one word in an image. For this, we performed segmentation to obtain one word in an image. The working of text recognition process is explained in Figure 6. A binarization algorithm was applied, which extracted the text from the noisy and shadow affected images.

**Figure 6.** A flowchart of the text recognition process.

4. Commonly Used Deep Learning Techniques

Deep learning is a technique of machine learning, which uses multiple layers of neural networks for classifying patterns [51]. Deep learning models use neural networks referred to as deep neural networks (DNNs). Simple neural networks only use two to three hidden layers, but in DNNs, hidden layers can be more than a hundred (deep mean numbers of hidden layers in network). A simple neural network takes the input at the input layer, and processes this input through hidden layers. Each neuron in the hidden layer is fully connected to all previous layer neurons. In a layer, neurons do not share their information; they operate independently from other neurons [52]. The last layer of the network will be fully connected. These fully connected neural networks cannot scale to larger images. Convolutional neural networks (CNNs) solve this problem; each neuron in the hidden layer is not fully connected to all previous layer neurons [53]. Only a few neurons of the previous layer are connected to each neuron. The shortcoming of CNNs is that they are unable to deal with variable size input and output; they only take fixed size input and generate fixed size output [54]. CNNs are not suitable for some applications such as speech and video processing. The algorithm for data preprocessing in our proposed method is described in Algorithm 1.

Algorithm 1 Proposed algorithm used for Data Preprocessing

```

1: procedure PRE-PROCESSING(Dataset)
2:   for Each image in dataset do
3:     ImageBackground = PreBinarization(Image)
4:     ImageGradient = CalculateGradient(ImageBackground)
5:     ImageSmoothen = SmoothenGradient(ImageGradient)
6:     ThresholdedImage = ThresholdedImage(ImageSmoothen)
7:     DilatedImage = ApplyDilation(ThresholdedImage)
8:     OutliersRemovedImages = ApplyConnectedComponentAnalysis
9:     LocalizedTextImages = ApplyTextLocalization
10:  return LocalizedTextImages

```

Recurrent neural networks (RNNs) are designed to connect the previous information with the present task. They tackle the issue of fixed size input and output [55]. Simple RNNs cannot handle long dependencies, and are only useful when the gap between past information and the needed place is small. Simple RNNs face the problem of a vanishing gradient and exploding. Long short-term memory (LSTM) networks are a special kind of RNN that deal with this problem by having cell memory using linear and logistic units [56]. An LSTM network consists of three gates: write gate, keep the gate and read gate. Write gate writes the information on the cell, keep gate will keep the information on the cell and read gate allows one to read the information from the gate.

In deep learning, attention mechanisms attracted a lot of interest. Initially, an attention mechanism was applied on sequence for sequence models in neural machine translation, but now they are used in many problems, such as image captioning and others [57]. In attention-based LSTM, the system gives a score or attention weight to each output of LSTM. Based on these attention weights, the system will predict the output with high accuracy.

Overview of RNN

In 1960, the first RNNs were introduced. Due to their contextual modeling ability, they became more well known. RNNs are especially popular to tackle the time series problem with different patterns. Basically, with recurrent layers, an RNN is a simple multi-layer perceptron (MLP). Suppose an RNN receives an input sequence with B input units, H

hidden units, and O output units. The values of activation's f_h and hidden units e_h of each layer can be calculated with these equations:

$$e_h(t) = \sum_{m=1}^M v_{mh} a_m(t) + \sum_{h=1}^H v_{hh'} f_{h'}(t-1) \quad (8)$$

$$f_{ht}(t) = \Theta_h(e_h(t)) \quad (9)$$

Here, $a_m(t)$ is the value of input m in time t , $e_n(t)$ and $f_n(t)$ are the input and the activation of the network to unit n , respectively, at time t . From unit m to unit n , the connection is represented as v_{mn} . Θ_h represents the activation function of hidden unit h . For the first time, RNNs were used for speech recognition by Robison [20]. For handwritten recognition, RNNs were used by Lee and Kim [40]. An input image as in Figures 7 and 8 was preprocessed to obtain the background and locate the text regions. An image background can be estimated by the pre binarization process. Firstly, a color image was converted into a gray scale image and then further converted into black and white images. Vertical profile projections were used to obtain the segmentation of words. Figures 9–11 are the visual representations of text.



Figure 7. Original image.



Figure 8. Binary image.



Figure 9. Cropped first word.



Figure 10. Cropped second word.



Figure 11. Cropped third word.

A bidirectional RNN was presented by Schuter and Paliwal [18] in 1997. There are two recurrent layers in sequence processing, one for the forward direction and the second for the backward direction. These recurrent layers are connected to the same input and output layers. To deal with multi-dimensions such as 2D images or 3D videos, Graves [41]

presents a multi-dimension recurrent neural network (MD-RNN). For multi-dimension RNN, consider a point of input series designed to have higher-layer units such as memory cells. These memory cells are specially used to maintain information for a longer time span.

In the 1D case, the input sequence is written as $e(t)$, but here, e^q is written as input for multi-dimension cases. The position of the upper index is written as $q_m, m \in 1,2,3, \dots k$. In dimension g , the step back positions can be presented as $Q'_g = q_1 \dots q_{d-1}, \dots q_k$. From m to n with dimension g , V_{mn}^d is the recurrent connection. The equation with the forward direction of g dimension MD-RNNs is represented as:

$$e_h^q = \sum_{m=1}^M v m h a_m^q + \sum_{g=1}^k \sum_{h'=1}^H f_h^{q-d} v_{h'}^q h \tag{10}$$

$$f_h^q = \Theta_h(e_h^q) \tag{11}$$

The below equation calculates the backward pass as follows:

$$e_h^q = \sum_{o=1}^O v h o \delta_m^q + \sum_{g=1}^k \sum_{h'=1}^H \delta_h^{q+d} v_{h'}^q h \tag{12}$$

$$\delta_h^q = \Theta_h(e_h^q) \in_h^q \tag{13}$$

where $e_n^q = \partial E / \partial f_n^q$ is the output error of n unit at time q and $\delta_n^q = \partial E / \partial e_n$ represents the error after accumulation.

One-dimensional recurrence is used by standard RNNs such as one axis of an image, but the scanning of an image by MDRNNs is carried out along both axes. This will enable more contextual variations in four directions (top, bottom, left, and right). Figure 12 illustrates the scanning direction of 2D-RNN. In the figure, a^q is the input sequence; the two points $(m-1, n)$ and $(m, n-1)$ always reached the end before that point (m, n) . There were very few practical applications of RNNs due to the problem of vanishing gradient and long-term dependencies. In 1997, Hochreiter designed a new network, LSTM, to tackle long-term dependencies. LSTMs are a special type of RNN that are designed to have hidden-layer units such as memory cells. These memory cells are specially used to maintain information for a longer time span.

$$i g_t = \sigma(W_{i g a} a_t + W_{i g h} h_{t-1} + W_{i g m c} m c_{t-1} + b i_{i g}) \tag{14}$$

$$f g_t = \sigma(W_{f g a} a_t + W_{f g h} h_{t-1} + W_{f g m c} m c_{t-1} + b i_{m c}) \tag{15}$$

$$m c_t = f g_t m c_{t-1} i g_t \tanh(W_{a m c} a_t + W_{h m c} h_{t-1} + W_{g h m c} m c_{t-1} + b i_{m c}) \tag{16}$$

$$o g_t = \sigma(W_{o g a} a_t + W_{o g h} h_{t-1} + W_{o g m c} m c_{t-1} + b i_{o g}) \tag{17}$$

$$h_t = o g_t \tanh(m c_t) \tag{18}$$

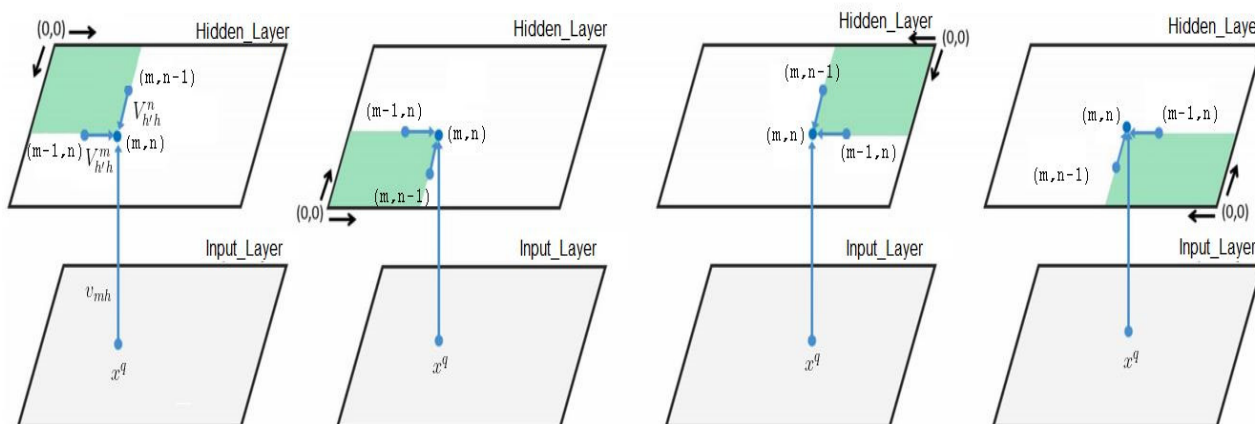


Figure 12. The scanning direction of 2D-RNN.

5. Description of Implementation

In the proposed method, the convolutional layer follows the VGG architecture [6]. In VGG architecture, square pooling windows are used, but this architecture uses rectangular max-pooling windows. This convolutional layer generates wider and longer feature maps. These feature maps will be the input of RNN layers. An illustration of LSTM architecture is given in Figure 13. Figure 14 compares proposed architecture with different architectures on character recognition rate using Alif dataset. However, for line recognition rate and word recognition rate, Figures 15 and 16 describe analytically. With respect to Activ dataset, the comparison of proposed architecture with different architecture on WRR, LRR and CRR are illustrated in Figures 17–19 respectively.

All the scene text and video text images are resized to a fixed height and width. The input resized shape for video text is (90,300) and for a scene, the text is (110,320). We have tested this many times, and observe that the performance is not much affected by resizing the input image into the fixed width and height. As is well known, the Arabic language starts from right to left, for this, all the input images are horizontally flipped; then, these will be the input to the convolutional layer. Here, the architecture composes three parts: the convolutional part, RNN part, and attention over RNN; because of them, it will face a training problem. The batch normalization technique is used to handle this problem [27]. Stochastic gradient descent is used to train this architecture. Backpropagation algorithms are used to calculate gradients. In [28], the backpropagation through time algorithm is used for differentials of error for RNN layers. The adadelata optimization technique is used to set the learning rate parameter [29].

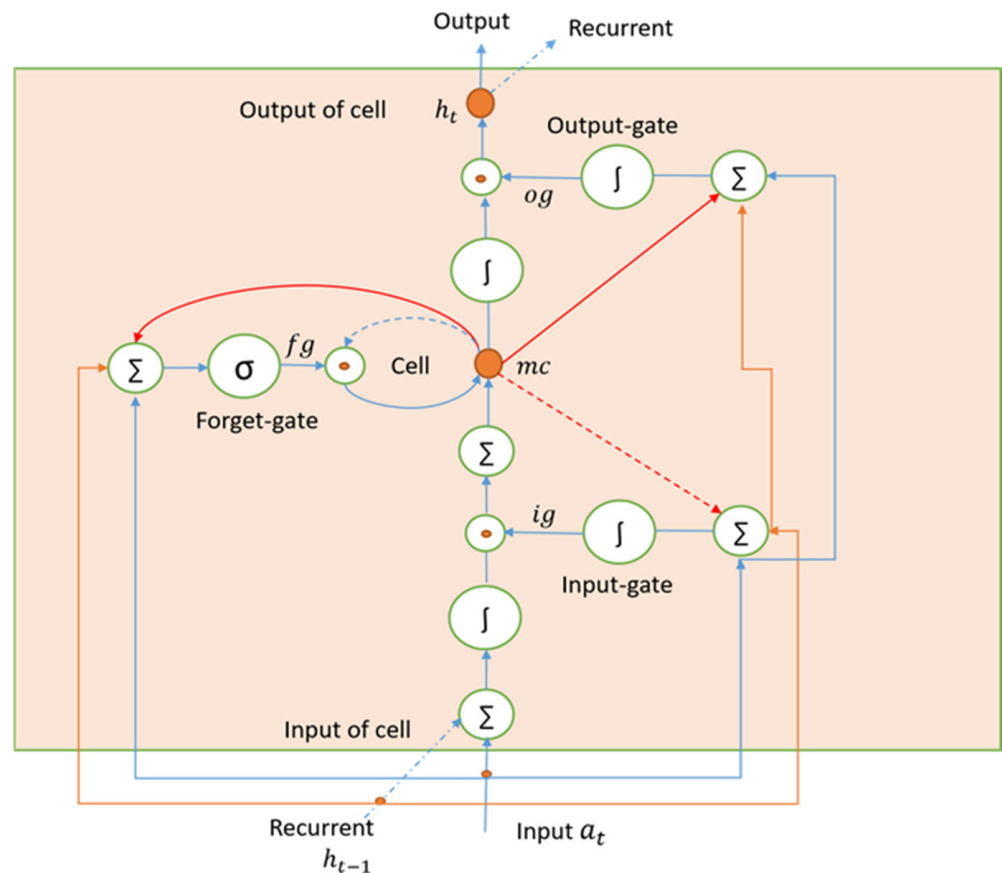


Figure 13. LSTM architecture.

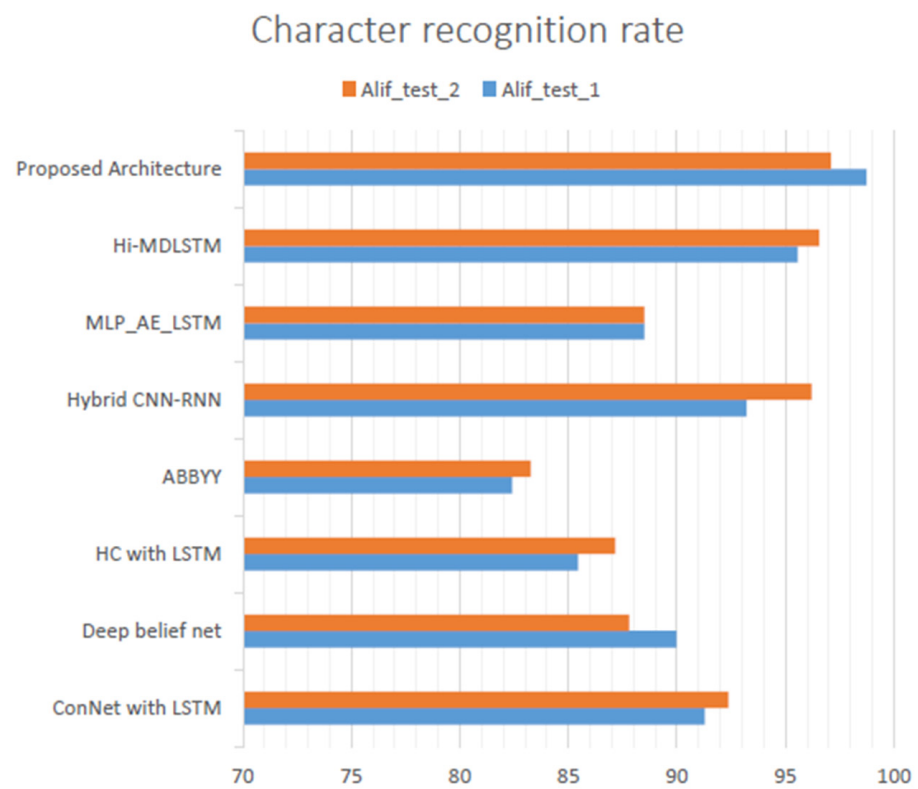


Figure 14. Comparison of proposed architecture with different architectures on character recognition rate (Alif dataset).

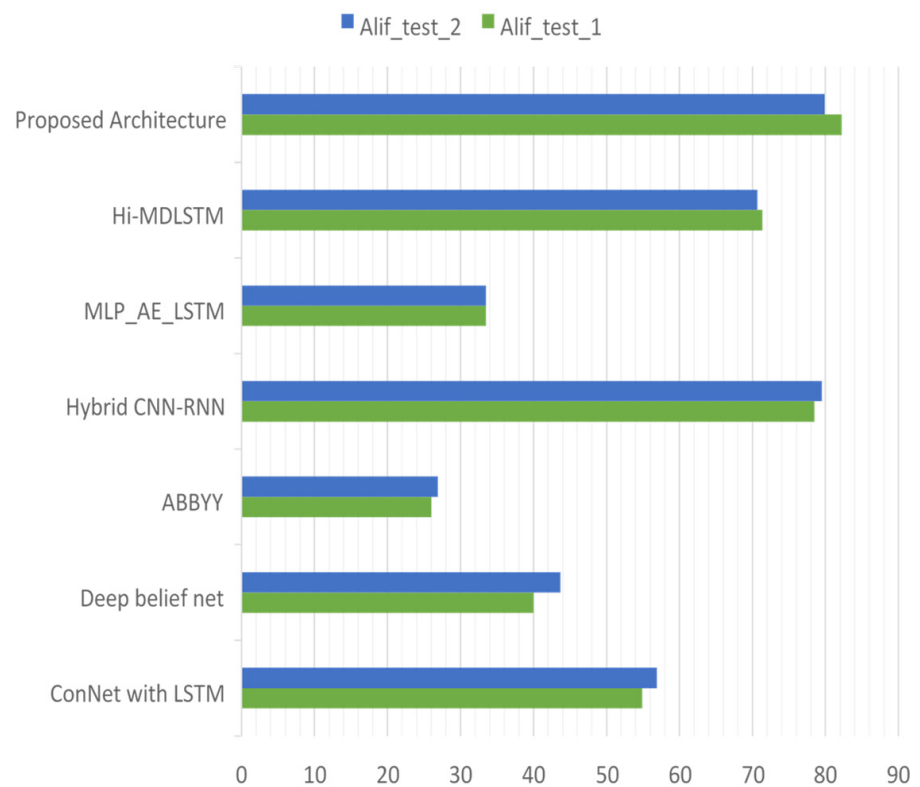


Figure 15. Comparison of proposed architecture with different architectures on line recognition rate (Alif dataset).

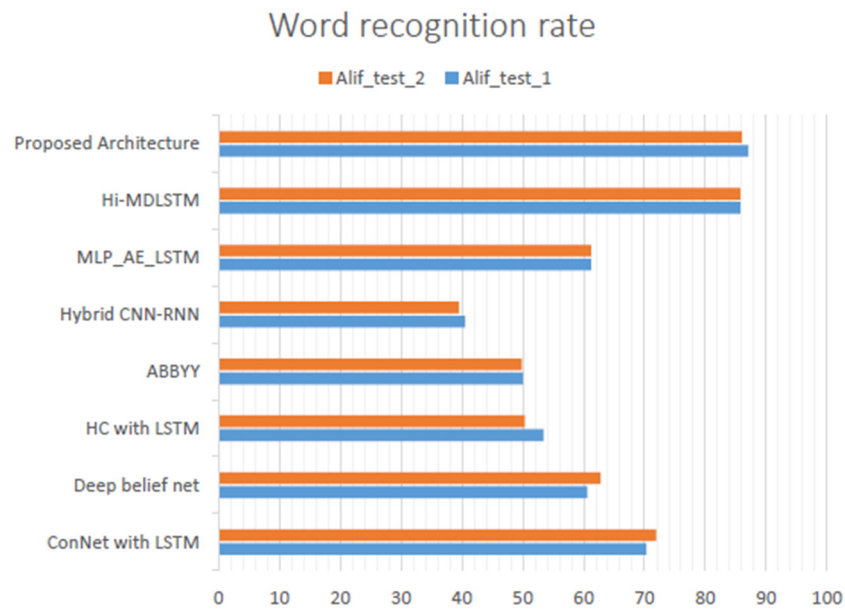


Figure 16. Comparison of proposed architecture with different architectures on word recognition rate (Alif dataset).

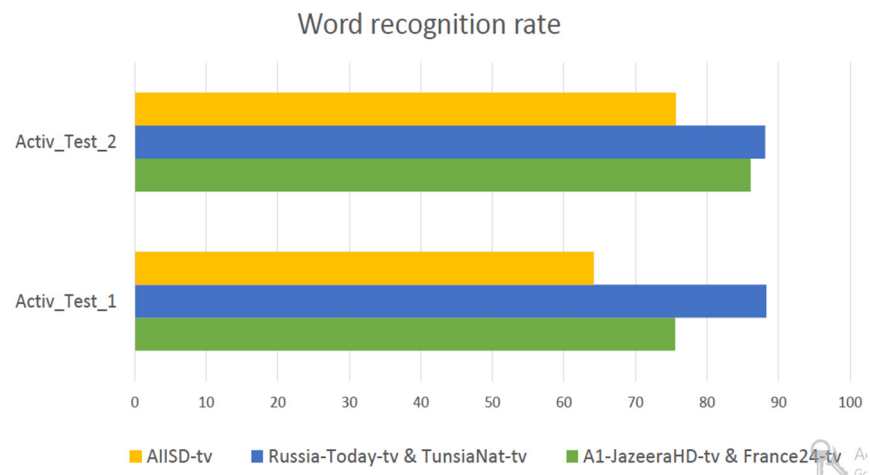


Figure 17. Comparison of proposed architecture with different architectures on word recognition rate (Activ dataset).

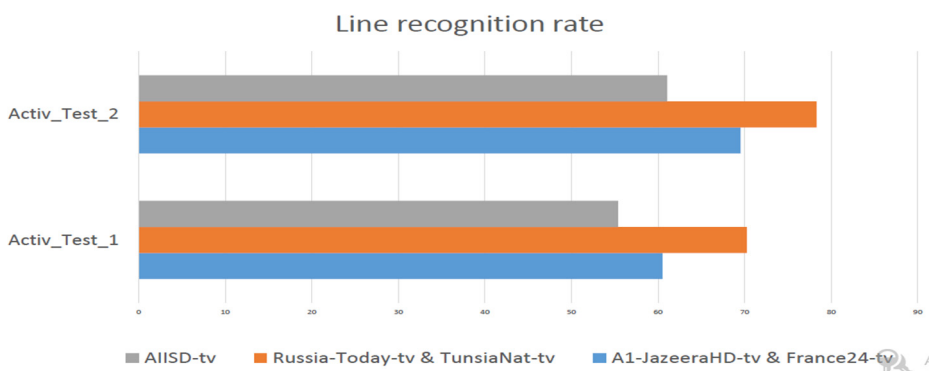


Figure 18. Comparison of proposed architecture with different architectures on line recognition rate (Activ dataset).

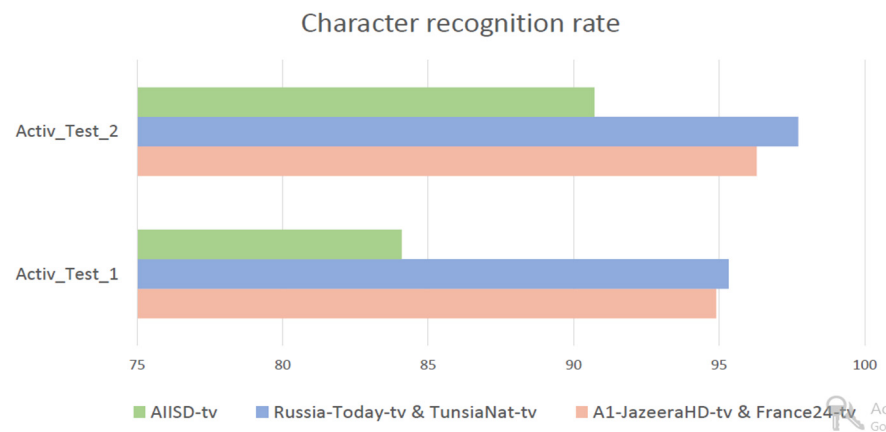


Figure 19. Comparison of proposed architecture with different architectures on character recognition rate (Activ dataset).

6. Experimental Results

The text on natural scene images and video frames can be recognized by the above-discussed architecture; this section will elaborate on the efficient recognition of Arabic text by this architecture. Here, as we know, all the input images are cropped from the original scene image or a video frame. Each image contains only one word of Arabic text. As per our knowledge, most of the work on Arabic text is character-level classification or recognition. This paper introduces new Arabic scene text baseline results. There are two datasets available for Arabic video text, Alif and Activ [23,25]; the results are reported on these two datasets. Table 4 presents the results of recognition on the Alif dataset for video text. Table 5 shows the proposed architectural results of the AcTiV dataset. A famous Optical character recognition Tesseract (OCR-T) [47] is applied over our one-word dataset for comparison, because there is no or a lower amount of work on a word-level dataset. Three metrics are used to evaluate the performance: Character Recognition Rate (Ch_{RR}), Word Recognition Rate (Wo_{RR}), and Line Recognition Rate (Li_{RR}).

$$Ch_{RR} = \frac{n_{chars} - \sum Edit_{dist}(R_{text}, G_{truth})}{n_{chars}} \quad (19)$$

$$Wo_{RR} = \frac{n_{words} - correctly - recognized}{n_{words}} \quad (20)$$

$$Li_{RR} = \frac{n_{textimages} - correctly - recognized}{n_{textimages}} \quad (21)$$

The methods used for comparison on video text recognition have different convolutional architecture for the extraction of features from end-to-end trainable architecture. For the video text recognition task, we obtained better character-level and line-level accuracy and set the new state of the art. A Deep Belief Network (DBN) is a feed-forward neural network with one or more layers of hidden units, which are commonly referred to as feature detectors [59]. The fact that generative weights can be learned layer-by-layer is a unique feature of DBN. A pair of layers' latent variables can be learned at the same time. MLP_AE_LSTM is an Auto Encoder Long Short-Term Memory based on Multi-Layer Perceptron [59]. This is a feed-forward network that learns to provide the same output as the input. In this network, they used a three-layered neural network with fully connected hidden units. The ability to collect multi-model features is improved by using more than one hidden layer with nonlinear units in the auto-encoder. The backpropagation algorithm is used to train the network. ABBYY [59] is a famous Arabic OCR engine Fine reader 12, used for Arabic character recognition; this OCR obtains a 82.4% to 83.26% character recognition rate, but our proposed model obtains a 98.73% character recognition rate.

MDLSTM [60] is a Multi-Dimension Long Short-Term Memory Network; here, MDLSTM was used for Arabic text recognition, and the text variations on both dimensions of the input image are modeled using this architecture.

Table 4. Results on Alif dataset and comparison with other architectures.

Architecture	Alif_Test_1			Alif_Test_2		
	<i>Ch_{RR}</i> (%)	<i>Li_{RR}</i> (%)	<i>Wo_{RR}</i> (%)	<i>Ch_{RR}</i> (%)	<i>Li_{RR}</i> (%)	<i>Wo_{RR}</i> (%)
ConNet with LSTM	91.27	54.9	70.29	92.37	56.9	71.9
Deep belief net	89.98	40.05	60.58	87.8	43.7	62.78
HC with LSTM	85.44	60.15	53.4	87.14	62.30	50.31
ABBY	82.4	25.99	50.0	83.26	26.91	49.80
Hybrid CNN-RNN	93.2	78.5	40.5	96.2	79.5	39.5
MLP_AE_LSTM	88.50	33.5	61.22	88.50	33.5	61.22
Hi-MDLSTM	95.55	71.33	85.72	96.55	70.67	85.71
Proposed Architecture	98.73	82.21	87.06	97.09	79.91	85.98

Table 5. Proposed architectural results on AcTiV dataset.

TV Channels	AcTiV_Test_1			AcTiV_Test_2		
	<i>Ch_{RR}</i> (%)	<i>Li_{RR}</i> (%)	<i>Wo_{RR}</i> (%)	<i>Ch_{RR}</i> (%)	<i>Li_{RR}</i> (%)	<i>Wo_{RR}</i> (%)
A1-JazeeraHD-tv and France24-tv	94.9	60.51	75.55	96.29	69.54	86.11
Russia-Today-tv and TunisiaNat-tv	95.33	70.28	88.31	97.72	78.33	88.14
AIISD-tv	84.09	55.39	64.17	90.71	61.07	75.64

7. Conclusions

This paper presents a state-of-the-art deep learning attention mechanism over an RNN which gives successful results over Arabic text datasets such as Alif and Activ. In recent works, RNN's performance in sequence learning approaches has been influential, especially in text transcription and speech recognition. The attention layer over RNN enables one to obtain a focused area of the input sequence and allows fast and easier learning. In preprocessing, we made a new dataset of one word on an image from Alif and Activ, which will help to improve the reduction in line error rate. We reported it as 85% to 87%. This model gives better results from those based on a simple CNN, RNN, and hybrid CNN-RNN. This model outperformed results with any language modeling and misfit regularization techniques. Images can be transcribed directly from sequence learning techniques and the context of them can be modeled in both directions, forward and backward. Due to the availability of good feature learning algorithms, the focus of the community should move to harder problems such as natural scene recognition.

Author Contributions: Conceptualization, H.B., M.J.A. and M.H.; methodology, M.R.R. and M.J.R.; formal analysis, M.R.R., M.J.R. and H.B.; investigation, M.H. and M.J.A.; resources, H.B. and M.R.R.; writing—original draft preparation, H.B., M.R.R., M.J.R. and M.H.; writing—review and editing, M.J.A., M.J.R., H.B. and M.R.R.; supervision, M.R.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: We have prepared dataset from two publicly available datasets Elif and Active which are publicly available. Data Citations Jain, M.; Mathew, M.; Jawahar, C.V. Unconstrained scene text and video text recognition for Arabic script. In Proceedings of the 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR); Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2017; pp. 26–30 and Oussama Zayene, A Dataset for Arabic Text Detection, Tracking and Recognition in News Videos—AcTiV (AcTiV), 1, ID:AcTiV_1, URL: http://tc11.cvc.uab.es/datasets/AcTiV_1 (accessed on 12 May 2021).

Acknowledgments: The authors would like to express their gratitude to the journal Forecasting for waiving the article’s APC.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lienhart, R.; Wernicke, A. Localizing and segmenting text in images and videos. *IEEE Trans. Circuits Syst. Video Technol.* **2002**, *12*, 256–268. [[CrossRef](#)]
2. Aldahiri, A.; Alrashed, B.; Hussain, W. Trends in Using IoT with Machine Learning in Health Prediction System. *Forecasting* **2021**, *3*, 181–206. [[CrossRef](#)]
3. Merler, M.; Galleguillos, C.; Belongie, S. Recognising groceries in situ using in vitro training data. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; IEEE: Piscataway Township, NJ, USA, 2007; pp. 1–8.
4. Bin Ahmed, S.; Naz, S.; Razzak, M.I.; Yusof, R. Arabic Cursive Text Recognition from Natural Scene Images. *Appl. Sci.* **2019**, *9*, 236. [[CrossRef](#)]
5. Hussain, W.; Hussain, F.K.; Saberi, M.; Hussain, O.K.; Chang, E. Comparing time series with machine learning-based prediction approaches for violation management in cloud SLAs. *Futur. Gener. Comput. Syst.* **2018**, *89*, 464–477. [[CrossRef](#)]
6. Saidane, Z.; Garcia, C. Automatic scene text recognition using a convolutional neural network. In *Workshop on Camera-Based Document Analysis and Recognition*; Imlab, September 2007; Volume 1. Available online: <http://www.m.cs.osakafu-u.ac.jp/cbdar2007/proceedings/papers/P6.pdf> (accessed on 12 May 2021).
7. Zayene, O.; Seuret, M.; Touj, S.M.; Hennebert, J.; Ingold, R.; Ben Amara, N.E. Text Detection in Arabic News Video Based on SWT Operator and Convolutional Auto-Encoders. In Proceedings of the 2016 12th IAPR Workshop on Document Analysis Systems (DAS), Santorini, Greece, 11–14 April 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2016; pp. 13–18.
8. De Campos, T.E.; Babu, B.R.; Varma, M. Character recognition in natural images. In Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, Lisboa, Portuga, 5 February 2009; Volume 7, pp. 273–280. [[CrossRef](#)]
9. Hussain, W.; Sohaib, O. Analysing Cloud QoS Prediction Approaches and Its Control Parameters: Considering Overall Accuracy and Freshness of a Dataset. *IEEE Access* **2019**, *7*, 82649–82671. [[CrossRef](#)]
10. Yi, C.; Tian, Y. Scene Text Recognition in Mobile Applications by Character Descriptor and Structure Configuration. *IEEE Trans. Image Process.* **2014**, *23*, 2972–2982. [[CrossRef](#)] [[PubMed](#)]
11. Wang, K.; Babenko, B.; Belongie, S. End-to-end scene text recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2011; pp. 1457–1464.
12. Gur, E.; Zelavsky, Z. Retrieval of Rashi Semi-cursive Handwriting via Fuzzy Logic. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2012; pp. 354–359.
13. Raza, M.R.; Varol, A. QoS Parameters for Viable SLA in Cloud. In Proceedings of the 2020 8th International Symposium on Digital Forensics and Security (ISDFS), Beirut, Lebanon, 1–2 June 2020; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2020; pp. 1–5.
14. Ahmed, S.; Pasquier, M.; Qadah, G.Z. Key issues in conducting sentiment analysis on Arabic social media text. In Proceedings of the 2013 9th International Conference on Innovations in Information Technology (IIT), Al Ain, United Arab Emirates, 17–19 March 2013; Institute of Electrical and Electronics Engineers (IEEE): Al Ain, United Arab Emirates, 2013; pp. 72–77.
15. Alma’ Adeed, S.; Higgins, C.; Elliman, D. Off-line Recognition of Handwritten Arabic Words Using Multiple Hidden Markov Models. In *Research and Development in Intelligent Systems XX*; Springer Science and Business Media LLC: Berlin, Germany, 2004; pp. 33–40.
16. Lakhfif, A.; Laskri, M.T. A frame-based approach for capturing semantics from Arabic text for text-to-sign language MT. *Int. J. Speech Technol.* **2015**, *19*, 203–228. [[CrossRef](#)]
17. Hussain, W.; Hussain, F.K.; Hussain, O. Maintaining Trust in Cloud Computing through SLA Monitoring. In Proceedings of the International Conference on Neural Information Processing, Kuching, Malaysia, 3–6 November 2014; pp. 690–697.

18. Jain, M.; Mathew, M.; Jawahar, C.V. Unconstrained scene text and video text recognition for Arabic script. In Proceedings of the 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, 3–5 April 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2017; pp. 26–30.
19. Chowdhury, A.; Vig, L. An efficient end-to-end neural model for handwritten text recognition. *arXiv* **2018**, arXiv:1807.07965.
20. Yousefi, M.R.; Soheili, M.R.; Breuel, T.M.; Stricker, D. A comparison of 1D and 2D LSTM architectures for the recognition of handwritten Arabic. In *Document Recognition and Retrieval XXII*; International Society for Optics and Photonics: Bellingham, WA, USA, 2015; Volume 9402. [[CrossRef](#)]
21. Chen, D.; Odobez, J.-M.; Boulard, H. Text detection and recognition in images and video frames. *Pattern Recognition*. **2004**, *37*, 595–608. [[CrossRef](#)]
22. Hussain, W.; Hussain, F.K.; Hussain, O.K.; Damiani, E.; Chang, E. Formulating and managing viable SLAs in cloud computing from a small to medium service provider’s viewpoint: A state-of-the-art review. *Inf. Syst.* **2017**, *71*, 240–259. [[CrossRef](#)]
23. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
24. Zayene, O.; Hennebert, J.; Touj, S.M.; Ingold, R.; Amara, N.E.B. A dataset for arabic text detection, tracking and recognition in news videos-activ. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; IEEE: Piscataway Township, NJ, USA, 2015; pp. 996–1000.
25. Yousfi, S.; Berrani, S.-A.; Garcia, C. ALIF: A dataset for Arabic embedded text recognition in TV broadcast. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2015; pp. 1221–1225.
26. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [[CrossRef](#)]
27. Alkalbani, A.M.; Hussain, W. Cloud service discovery method: A framework for automatic derivation of cloud market-place and cloud intelligence to assist consumers in finding cloud services. *Int. J. Commun. Syst.* **2021**, *34*, e4780. [[CrossRef](#)]
28. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
29. Mikolov, T.; Karafiát, M.; Burget, L.; Cernock, Y.; Khudanpur, S. Recurrent neural network based language model. In Proceedings of the 11th Annual Conference of the International Speech Communication Association, Makuhari, Japan, 26–30 September 2010.
30. Shi, B.; Bai, X.; Yao, C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2298–2304. [[CrossRef](#)]
31. Alrashed, B.A.; Hussain, W. Managing SLA Violation in the cloud using Fuzzy re-SchedNeg Decision Model. In Proceedings of the 2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA), Kristiansand, Norway, 9–13 November 2020; IEEE: Piscataway Township, NJ, USA, 2020.
32. Graves, A.; Wayne, G.; Danihelka, I. Neural Turing machines. *arXiv* **2014**, arXiv:1410.5401.
33. Bissacco, A.; Cummins, M.; Netzer, Y.; Neven, H. PhotoOCR: Reading Text in Uncontrolled Conditions. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2013; pp. 785–792.
34. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attentionbased models for speech recognition. *arXiv* **2015**, arXiv:1506.07503.
35. Haddad, S.E.; Roitfarb, H.R. The structure of arabic language and orthography. In *Handbook of Arabic Literacy*; Springer: Berlin, Germany, 2014; pp. 3–28.
36. Gillies, A.; Erlandson, E.; Trenkle, J.; Schlosser, S. Arabic Text Recognition System. In Proceedings of the Symposium on Document Image Understanding Technology; 1999; pp. 253–260. Available online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.21.947&rep=rep1&type=pdf> (accessed on 12 May 2021).
37. Shahin, A.A. Printed Arabic Text Recognition using Linear and Nonlinear Regression. *Int. J. Adv. Comput. Sci. Appl.* **2017**, *8*. [[CrossRef](#)]
38. Halima, M.B.; Alimi, A.; Vila, A.F. Nf-savo: Neuro-fuzzy system for arabic video ocr. *arXiv* **2012**, arXiv:1211.2150.
39. Hussain, W.; Hussain, F.K.; Hussain, O.; Chang, E. Profile-Based Viable Service Level Agreement (SLA) Violation Prediction Model in the Cloud. In Proceedings of the 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), Krakow, Poland, 4–6 November 2015; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2015; pp. 268–272.
40. Iwata, S.; Ohyama, W.; Wakabayashi, T.; Kimura, F. Recognition and transition frame detection of Arabic news captions for video retrieval. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2016; pp. 4005–4010.
41. Alrehali, B.; Alsaedi, N.; Alahmadi, H.; Abid, N. Historical Arabic Manuscripts Text Recognition Using Convolutional Neural Network. In Proceedings of the 2020 6th Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 4–5 March 2020; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2020; pp. 37–42.

42. Younis, K.; Khateeb, A. Arabic Hand-Written Character Recognition Based on Deep Convolutional Neural Networks. *Jordanian J. Comput. Inf. Technol.* **2017**, *3*, 186. [[CrossRef](#)]
43. El-Sawy, A.; Loey, M.; Bakry, H.E. Arabic handwritten characters recognition using convolutional neural network. *WSEAS Trans. Comput. Res.* **2017**, *5*, 11–19.
44. Torki, M.; Hussein, M.E.; Elsallamy, A.; Fayyaz, M.; Yaser, S. Window-based descriptors for arabic handwritten alphabet recognition: A comparative study on a novel dataset. *arXiv* **2014**, arXiv:1411.3519.
45. Alkalbani, A.M.; Hussain, W.; Kim, J.Y. A Centralised Cloud Services Repository (CCSR) Framework for Optimal Cloud Service Advertisement Discovery From Heterogenous Web Portals. *IEEE Access* **2019**, *7*, 128213–128223. [[CrossRef](#)]
46. Ahmad, R.; Naz, S.; Afzal, M.Z.; Rashid, S.F.; Liwicki, M.; Dengel, A. A deep learning based arabic script recognition system: Benchmark on khat. *Int. Arab J. Inf. Technol.* **2020**, *17*, 299–305.
47. Mahmoud, S.A.; Ahmad, I.; Alshayeb, M.; Al-Khatib, W.G.; Parvez, M.T.; Fink, G.; Margner, V.; El Abed, H. KHATT: Arabic Offline Handwritten Text Database. In Proceedings of the 2012 International Conference on Frontiers in Handwriting Recognition, Bari, Italy, 18–20 September 2012; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2012; pp. 449–454.
48. Slimane, F.; Ingold, R.; Hennebert, J. ICDAR2017 Competition on Multi-Font and Multi-Size Digitally Represented Arabic Text. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2017; Volume 1, pp. 1466–1472.
49. Alghamdi, A.; Hussain, W.; Alharthi, A.; Almusheqah, A.B. The Need of an Optimal QoS Repository and Assessment Framework in Forming a Trusted Relationship in Cloud: A Systematic Review. In Proceedings of the 2017 IEEE 14th International Conference on e-Business Engineering (ICEBE), Shanghai, China, 4–6 November 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2017; pp. 301–306.
50. Zhai, C.; Chen, Z.; Li, J.; Xu, B. Chinese Image Text Recognition with BLSTM-CTC: A Segmentation-Free Method. In *Communications in Computer and Information Science*; Springer Science and Business Media LLC: Berlin, Germany, 2016; pp. 525–536.
51. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Reading Text in the Wild with Convolutional Neural Networks. *Int. J. Comput. Vis.* **2016**, *116*, 1–20. [[CrossRef](#)]
52. Almazan, J.; Gordo, A.; Fornes, A.; Valveny, E. Word Spotting and Recognition with Embedded Attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 2552–2566. [[CrossRef](#)] [[PubMed](#)]
53. Pinheiro, P.H.; Collobert, R. Recurrent convolutional neural networks for scene labeling. In Proceedings of the 31st International Conference on Machine Learning (ICML), Beijing, China, 21–26 June 2014.
54. Gatta, C.; Romero, A.; van de Weijer, J. Unrolling Loopy Top-Down Semantic Feedback in Convolutional Deep Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Columbus, OH, USA, 23–28 June 2014; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2014; pp. 504–511.
55. Byeon, W.; Breuel, T.M.; Raue, F.; Liwicki, M. Scene labeling with LSTM recurrent neural networks. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2015; pp. 3547–3555.
56. Karray, H.; Ellouze, M.; Alimi, A.M. Indexing Video Summaries for Quick Video Browsing. In *Computer Communications and Networks*; Springer Science and Business Media LLC: Berlin, Germany, 2009; pp. 77–95.
57. Hua, X.S.; Chen, X.-R.; Wenyin, L.; Zhang, H.-J. Automatic location of text in video frames. In Proceedings of the 2001 ACM Workshops on Multimedia: Multimedia Information Retrieval, Ottawa, ON, Canada, 30 September–5 October 2001; ACM: New York, NY, USA, 2001; pp. 24–27.
58. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In *Advances in Neural Information Processing Systems*; NIPS: Grenada, Spain, 2014; pp. 2204–2212.
59. Hussain, W.; Hussain, F.K.; Hussain, O. Comparative analysis of consumer profile-based methods to predict SLA violation. In Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul, Turkey, 2–5 August 2015; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2015; pp. 1–8.
60. Kim, S.; Hori, T.; Watanabe, S. Joint CTC-attention based end-to-end speech recognition using multi-task learning. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2017; pp. 4835–4839.
61. Gers, F.A.; Schraudolph, N.N.; Schmidhuber, J. Learning precise timing with LSTM recurrent networks. *J. Mach. Learn. Res.* **2002**, *3*, 115–143.
62. Graves, A.; Mohamed, A.-R.; Hinton, G. Speech recognition with deep recurrent neural networks. In Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–30 May 2013; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2013; pp. 6645–6649.
63. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [[CrossRef](#)]
64. Shea, K.O.; Nash, R. An introduction to convolutional neural networks. *arXiv* **2015**, arXiv:1511.08458.
65. Karsoliya, S. Approximating number of hidden layer neurons in multiple hidden layer BPNN architecture. *Int. J. Eng. Trends Technol.* **2012**, *3*, 714–717.

66. Hussain, W.; Hussain, F.; Hussain, O. QoS prediction methods to avoid SLA violation in post-interaction time phase. In Proceedings of the 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), Hefei, China, 5–7 June 2016; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2016; pp. 32–37.
67. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [[CrossRef](#)]
68. Sutskever, I.; Vinyals, O.; Le, Q.V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*; NIPS: Grenada, Spain, 2014; pp. 3104–3112.
69. Yao, K.; Cohn, T.; Vylomova, K.; Duh, K.; Dyer, C. Depth-gated lstm. *arXiv* **2015**, arXiv:1508.03790.
70. Luong, M.T.; Pham, H.; Manning, C.D. Effective Approaches to Attention-based Neural Machine Translation. *arXiv* **2015**, arXiv:1508.04025.
71. Hussain, W.; Sohaib, O.; Naderpour, M.; Gao, H. Cloud Marginal Resource Allocation: A Decision Support Model. *Mob. Netw. Appl.* **2020**, *25*, 1418–1433. [[CrossRef](#)]
72. Fasha, M.; Hammo, B.; Obeid, N.; Alwidian, J. A Hybrid Deep Learning Model for Arabic Text Recognition. *Int. J. Adv. Comput. Sci. Appl.* **2020**, *11*. [[CrossRef](#)]
73. Yousfi, S.; Berrani, S.-A.; Garcia, C. Deep learning and recurrent connectionist-based approaches for Arabic text recognition in videos. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; Institute of Electrical and Electronics Engineers (IEEE): Piscataway Township, NJ, USA, 2015; pp. 1026–1030.
74. GarciaGraves, A. Offline Arabic Handwriting Recognition with Multidimensional Recurrent Neural Networks. In *Guide to OCR for Arabic Scripts*; Springer Science and Business Media LLC: Berlin, Germany, 2012; pp. 297–313.
75. Ahmed, R.; Gogate, M.; Tahir, A.; Dashtipour, K.; Al-Tamimi, B.; Hawalah, A.; El-Affendi, M.A.; Hussain, A. Deep neural network-based contextual recognition of arabic handwritten scripts. *Entropy* **2021**, *23*, 340. [[CrossRef](#)] [[PubMed](#)]
76. Nurseitov, D.; Bostanbekov, K.; Alimova, A.; Abdallah, A.; Abdimanap, G. Classification of Handwritten Names of Cities and Handwritten Text Recognition using Various Deep Learning Models. *Adv. Sci. Technol. Eng. Syst. J.* **2020**, *5*, 934–943. [[CrossRef](#)]
77. Hussain, W.; Hussain, F.K.; Hussain, O.K. Towards Soft Computing Approaches for Formulating Viable Service Level Agreements in Cloud. In *Transactions on Petri Nets and Other Models of Concurrency XV*; Springer Science and Business Media LLC: Berlin, Germany, 2015; pp. 639–646.
78. Altwajry, N.; Turaiki, I.A. Arabic handwriting recognition system using convolutional neural network. *Neural Comput. Appl.* **2021**, *33*, 2249–2261. [[CrossRef](#)]
79. Mirza, A.; Siddiqi, I. Recognition of cursive video text using a deep learning framework. *IET Image Process.* **2020**, *14*, 3444–3455. [[CrossRef](#)]
80. El Bazi, I.; Laachfoubi, N. Arabic named entity recognition using deep learning approach. *Int. J. Electr. Comput. Eng.* **2019**, *9*, 2025–2032.
81. Arafat, S.Y.; Iqbal, M.J. Urdu-text detection and recognition in natural scene images using deep learning. *IEEE Access* **2020**, *8*, 96787–96803. [[CrossRef](#)]