Global Journal of Computer Science and Technology: D
Neural & Artificial Intelligence
Volume 19 Issue 1 Version 1.0 Year 2019
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

Sub-Sampling Approach for Unconstrained Arabic Scene Text Analysis by Implicit Segmentation based Deep Learning Classifier

By Saad Bin Ahmed, Zainab Malik, Muhammad Imran Razzak & Rubiyah Yusof

King Saud bin Abdulaziz University for Health Sciences

Abstract- The text extraction from the natural scene image is still a cumbersome task to perform. This paper presents a novel contribution and suggests the solution for cursive scene text analysis notably recognition of Arabic scene text appeared in the unconstrained environment. The hierarchical sub-sampling technique is adapted to investigate the potential through sub-sampling the window size of the given scene text sample. The deep learning architecture is presented by considering the complexity of the Arabic script. The conducted experiments present 96.81% accuracy at the character level. The comparison of the Arabic scene text with handwritten and printed data is outlined as well.

Keywords: *sub sampling, MDLSTM, deep learning, Implicit segmentation, unconstraint.*

GJCST-D Classification: *I.2.6*

Strictly as per the compliance and regulations of:

Sub-Sampling Approach for Unconstrained Arabic Scene Text Analysis by Implicit Segmentation based Deep Learning Classifier

Saad Bin Ahmed [α ρ], Zainab Malik [α], Muhammad Imran Razzak [σ] & Rubiyah Yusof [α]

Abstract- **The text extraction from the natural scene image is still a cumbersome task to perform. This paper presents a novel contribution and suggests the solution for cursive scene text analysis notably recognition of Arabic scene text appeared in the unconstrained environment. The hierarchical sub-sampling technique is adapted to investigate the potential through sub-sampling the window size of the given scene text sample. The deep learning architecture is presented by considering the complexity of the Arabic script. The conducted experiments present 96.81% accuracy at the character level. The comparison of the Arabic scene text with handwritten and printed data is outlined as well.**

Keywords: sub sampling, MDLSTM, deep learning, Implicit segmentation, unconstraint.

I. Introduction

The research on unconstrained scene text recognition is gaining momentum for few years. The text separation always been a cumbersome task because the presence of other objects in an image. Although text provides information and guide in a situation having strange environment. It is essential to investigate about nature of a text appeared in a scene image so that it may provide meaning for someone. But the unconstrained scripts like the Arabic poses a huge challenge to deal with the complexities of language itself in the presence of other image degrading properties. The normal way to tackle with the problem of Arabic scene text classification, we usually disintegrate the part of an image into smaller units and investigate each one individually. Each Arabic character has four variations concerning its position appeared in a word i.e., a character can appear in isolation, at first, middle or at last position in a word. To overcome these implicit challenges, there are numerous techniques proposed recently [1, 5, 7, 8], which presented various feature extraction or classification techniques.

The nature of unconstrained Arabic script prompt researchers to suggest implicit segmentation approaches to deal with the complexity of under discussion script. To deal with the representation of the same character appears to be extreme difficult task to address. In this way, manual segmentation also proves to be a laborious work. We are looking for such type of solutions which proved good results on cursive scripts. This particular complexity of Arabic script prompts to suggest implicit segmentation techniques. The other important aspect of cursive scripts is to consider the context. In Arabic every character appearance depends on the previous character, in this way learning the context of current character is crucial. There are some solutions suggested by recent research to tackle with the variability of characters with context learning approaches as proposed in [9, 10, 11]. The most prominent context learning approach specifically used for unconstrained cursive text research is Long Short Term Memory (LSTM) networks [4].

By keeping in view the complexity associated with the cursive script, it is assumed that if scene image disintegrates into smaller parts then consider their feature values individually and assemble them together in one unit before applying the language model. For the cursive script like Arabic we require more detailed features of given patterns so that we may scrutinize and learn the pattern. Therefore, there is a need to look for such classification model which does not only learns the patterns from right to left or left to right but also from top to bottom and bottom to top. To address the problem above, this paper is proposing an adapted Multidimensional Long Short Term Memory (MDLSTM) networks [12]. The implicit segmentation approaches are more accurate and less error prone in comparison to those approaches defined explicitly. The parts of a given image are considered by the convolutional neural network (ConvNets) using implicit segmentation approach. As nature of ConvNets make it as instance learner, but there is a need to learn the context of a given sample in this way history of learned pattern play a role. Therefore, this paper is proposing deep learn- ing MDLSTM network because of its strong ability to learn sequence-based on the context. The Connectionist Temporal Classification (CTC) is used as a probabilistic model to map the learned sequences against corresponding ground truth [13]. By using CTC, explicit segmentation and modeling language is avoided. The performance of proposed MDLSTM network architecture is evaluated on Arabic scene text images. The EASTR-

Author α: Center of Aritifical Intelligence and Robotics (CAIRO-ikhoza), MJIIT, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia.
Author σ: University of Technology, Sydney, Australia.
Author ρ: King Saud bin Abdulaziz University for Health Sciences, Riyadh, Saudi Arabia. e-mails: saad2@utm.my, ahmedsa@ksau-hs.edu.sa, imran.razzak@ieee.org, rubiyah.kl@utm.my

42k dataset used for proposed work which covers various aspects of scene text images. The dataset contains 14, 000 segmented Arabic scene text images.

Arabic like languages share the same writing style i.e., from right to left. Arabic like scripts categorized into two forms i.e., joiner and non-joiner. The characters that appear as a joiner may join predecessor or successor character in a word, its mean these characters can appear as first, middle or at final position in a word. Whereas, non-joiner characters may appear in isolation or as the last character in a ligature. As mentioned earlier every character has option to appear on any of the four locations i.e., initial, middle, final or an isolated position. As far as Arabic scene text is concerned, it is relatively difficult to deal with the complexity of joined and non-joined characters. In camera captured text images, there are other numerous factors to concentrate on so that we may extract the text with high precision. There are numerous factors like illumination, an angle of a text, font size, appearance and clarity of a text pose a challenge for researchers to recognize Arabic scene text.

II. Related Work

There are various feature extraction and classification approaches has been proposed for detection and recognition of Latin and cursive scripts like Chinese in natural scene images [6, 14, 15]. The text in natural scene images not only represent the pattern information but also it exhibits semantic information which shows some meaning in real applications. This paper is presenting unconstrained character recognition in natural images having Arabic text in focus. By reviewing recent year's research, there is an impression that not enough work has been presented on Arabic text recognition in natural scene images. Although some substantial work have been reported which we summarized in this section.

One of the recent work on Arabic scene text is represented by [16]. They proposed Convolutional Recurrent Neural Network (CRNN) approach to evaluate the performance of their own gathered dataset and two publicly available video text datasets i.e., ALIF [17] and ACTIV [18]. They gathered 500 Arabic word images appeared in natural images. They categorized their experiments into character, word and line recognition. They reported very good accuracy on screen rendered video text datasets. The achieved 98.17% on character recognition, 79.67% accuracy on word recognition and 67.08% line recognition accuracy while on their own gathered dataset they achieved 69.55% and 39.25% accuracy on character and word recognition respectively. Another paper on deep learning based isolated Arabic scene character recognition is presented by [7]. They proposed deep convolutional neural net-work (ConvNets) architecture for recognition of Arabic characters appears in natural scene images. The features extraction and classification were per- formed through ConvNets. As there is not any benchmark dataset available for Arabic text in natural images they prepared dataset by their own which covers approximately every variation of each character. The experimental settings were empirically adjusted on 3

3 and 5 5 filter size with learning rate 0.5 and 0.005 by keeping the stride value 1 and 2. They identified 27 classes and save each character image with five orientations in different angles. In this way they identified 2450 images as the train set while 250 character images were used to evaluate the performance of their proposed algorithm. They reported 0.15% error rate on their proposed architecture.

The Arabic scene text dataset is proposed by [3]. They collected free Arabic text appeared in an unconstrained environment. They clicked 364 images having Arabic text. The images were segmented into 1280 cropped words. They also segmented acquired Arabic text into 374 characters. The major drawback of their proposed dataset is lack of applicability details.

The recent work on recognition and establishing a connection of moving Arabic text appeared in the video is presented by [19]. They developed dedicated OCR for the purpose to recognize low-resolution news captions in video images. They prepared dataset from Aljazeera news programs. They used connection method approach based on insertion operation, voting processing and substitution using minimum likelihood edit distance between two successive news frames for the purpose to connect text. Their proposed method is for automatic language translation and also helpful in reducing OCR errors caused by truncated characters. Their dataset was disintegrated into the train and the test set by using 453 video frames. They reported 96.78% accuracy through f-measure using bi-gram sequence.

III. Learning Architecture

a) Feature Extraction by ConvNets

The natural images have characteristics of representing the image details at the same level, meaning that representation of text in a natural image would seems at the same energy level as the other objects in an image. As we are dealing with text specifically so we were looking such a technique by which we can focus on text only in a natural image. The arbitrary size of

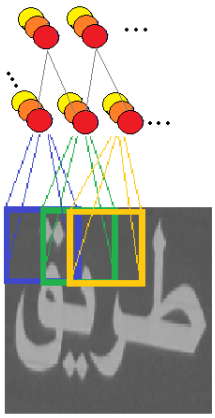

Figure 1: Convolutional Feature extraction

an input image is taken into account and normalized it with fixed 150 150 pixel size by considering the aspect ratio according to the image size. After that image is converted into gray scale. The 8 8 window is used to detect features of an image and make a feature map. This helps in considering each part of an image and focused on most relevant features of the corresponding image. At each point where the feature detector stops it takes a mean of involved pixels and write it over feature map at (1, 1) position. For the next move, the feature detector will move one pixel right and perform the same process again until the end of first row. After operating on first row feature detector window will move one step down to the second row and start the same process. In this way whole image will be filtered through feature detector window and update in feature map. The feature map contains a large amount of features in relevance to single image. Let's assume a small patch x x y, then array of convolved pixels will be represented as,

$$f_{convolve} = \eta(r - x + 1) * (c - y + 1) \quad (1)$$

The features f are obtained by taking mean η of contributed pixels r, c appears in feature detector window. The feature values write on the feature map from (1,1)(1,2)(1,3)....(143,143). Further explains the idea that feature map is mapping 143 features computed by applying mean pooling strategy. There considered 143 feature points corresponding to the given image. These extracted features are now ready to pass them to classifier.

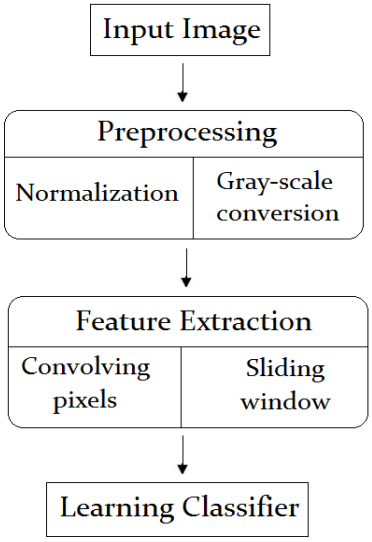

Figure 2: The flowchart of proposed idea

b) Multidimensional LSTM classifier for Arabic Scene text

The LSTM has effectively applied on a number of problems where data is correlated and sequence is important to learn. The correlation of data may be represented by single or multidimensional axis. The LSTM is a technique under RNN approach where unlike RNN the data can be modeled into multidimensional vector in addition to the single axis. The Arabic script recognition is a classic example of sequence learning tasks where context is important to learn. The representation of each character depends on the previous character and so on. Unlike Latin, the Arabic script written in joining style which complicate the recognition process. The ConvNets can prove to be the good choice to learn the different segments of handwritten samples which require a lot of manual pre-processing. Moreover, it cannot produce good results when the problem is large and where context learning is important.

The idea of multidimensional LSTM is to replace the single memory block of LSTM with the number of memory blocks according to dimensions. The input is delegated to hidden layers where the input data is processed by LSTM memory blocks in each dimension. In MDLSTM the self-connection of LSTM cell is controlled by n self-connection with n dimension and n forget gates. The cell activation values were forward to gates by peephole connections. The input gate in a memory block connected to all previous cells and in all dimensions. This will help to learn the sequential pattern of learning. The forget gate connected to cell c of all dimensions with different weights. This helps in determining how much previous computation takes part in all dimensions with reference to the current cell's computation. This type of setup is very important for Arabic script recognition where each character has four

variations according to the position in a word, moreover the character segmentation is also extremely difficult.

The MDLSTM is considered as ideal architecture for learning the sequential problems more efficiently and effectively. Most of the recently reported work on Urdu and in Arabic script recognition as explained in [2] proposed MDLSTM for learning the complex patterns and reported state-of-the-art results. The details about MDLSTM network architecture can be explored in Graves et al. paper [12].

IV. Hierarchical Subsampling based Cursive Document and Scene Text Recognition

The adapted hierarchical MDLSTM architecture based on sub-sampling of hidden layers approach is proposed for Arabic scene text.

The hierarchical subsampling usually applies where the data volume is too large and complex. The hierarchical subsampling based LSTM architecture includes input layer, an output layer and multiple self-connected hidden layers. The output of each level in the hierarchy is represented as input to the level up and so on. The input sequences were subsampled by predetermined window width. The hierarchical subsampling of RNN based networks follows the same structure as defined for ConvNets. The potential of sub-sampling approach was scrutinized by investigating the performance through 3 layer architecture which incorporates 20, 40, 60, 80, 100 and 120 hidden memory block sizes.

Table 1: Selected Parameters during training the network

Parameters	Values
Input block size	4×1
Hidden block size	4
Subsample sizes	6 and 20
Hidden sizes	2,10 and 50
Learn rate	1×10^{-4}, 1×10^{-3}
Momentum	0.9
Total network weight	732863

The network learning is based on the empirically selected parameters. The prime objective is to look for appropriate parameters that provide low error rate in comparison. The parameters detail along error rates and overall training time is provided in Table 1. The Arabic word assorted from the scene text initially pre-process to the standard size of 70 by keeping the aspect ratio. The feature map is prepared by convolving the extracted features from given image through filter window. The convolution process

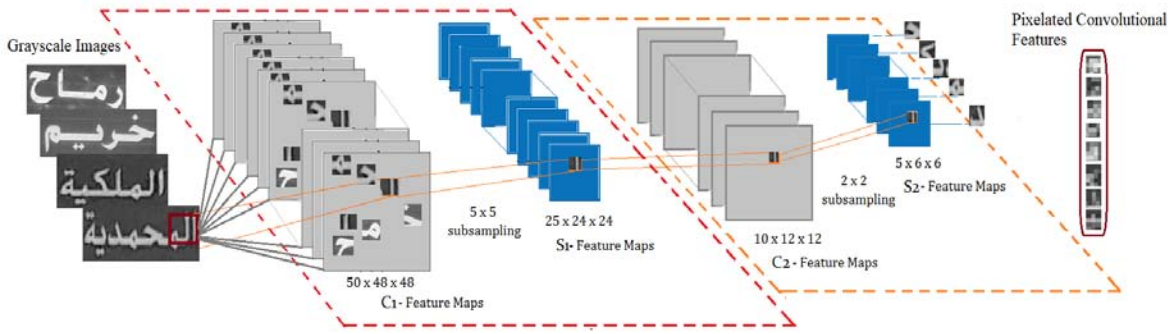

Figure 3: Arabic scene text feature extraction by convolutional pixelate method

is similar as presented in section ?? for handwritten Urdu text as depicted in Figure 3. Here, the gray scale values of convolved pixels are passed to classifier by following a specific input size as sketched in Figure 4.

In each feature map, every neuron is mapped according to small 5 $\times$ 5 region of an input image. The connection from input image to hidden layer is established through local receptive field called a filter size. Each neuron in a layer shares the same bias value. As single feature map does not cover the intensive features, therefore the process is further delegated to have a variety of features against each given image. A feature map is defined by its share weight and a bias value; mathematically this relation can be represented as follows in equation 2,

$$\alpha\left(d + \sum_{e=0}^{4}\sum_{f=0}^{4} W_{e,f} A_{j+e,k+f}\right) \quad (2)$$

whereas, α is neural activation sigmoid function while d is a shared value of bias. We,f represents filter or kernel weight which depends on filter size whereas, A represents the input activation at point (x,y).

The extracted features by ConvNets are converted into raw pixels and are given to MDLSTM architecture with corresponding ground truth as presented in Figure 4. The complex nature of Arabic script prompts to proposed a hierarchical subsampling architecture of MDLSTM for learning purpose. The proposed experiments are based on the subsampling architecture which is divided into two main categories. As a first evaluation, the experiments were performed having 3 and 5 layers architecture. Each layer incorporate 20, 40, 60, 80, 100 and 120 hidden LSTM memory block. The three-layer architecture is defined by number of hidden memory units at every three layers. The input is subsampled by 6 6 and 2 9 window size. The deep learning architecture is designed by defining the data into layer wise manner. The same process is applied on five layer architecture.

The second variation of experiments performed by defining the same pa- rameters as experimented by [12, 20]. [12] proposed their solution on hand- written Arabic character recognition while [20] presented the same idea on printed Urdu character recognition using similar parameters. The same pa- rameters and network structure are deliberately to compare the performance of handwritten, printed and scene text Arabic script recognition as shown in Figure 4. All activation functions in sub-sampling layers are feed forward tanh layers, whereas hidden layers are fully connected in all dimensions. The MDLSTM network collapse all processing into one dimensional CTC layer having 40 classes including a blank label which predict the output symbol. All activation functions in sub-sampling layers are feed forward tanh layers, whereas hidden layers are fully connected in all dimensions. The MDL-

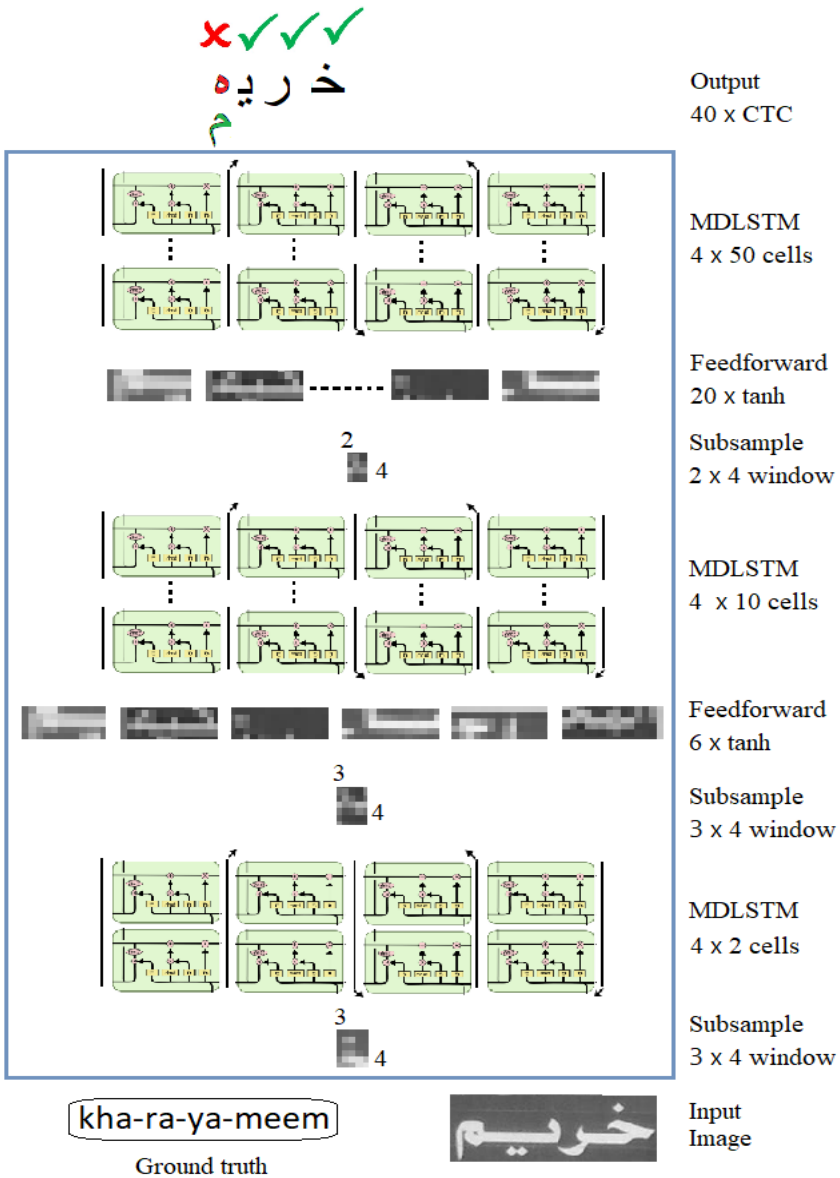

Figure 4: Hierarchical sub-sampling approach. As indicated in the output the character 'meem' (in green) not recognized by the network.

Table 2: Selected Parameters during training the network

Parameters	Values	Training/ Validation Error	No. of Epochs	Time/Epoch (minutes)
Subsample window	6×6	0.86/ 0.83	317	40
	2×9	0.94/ 0.92	299	34
Hidden memory units	20,60,100, 120	Best(0.97/0.95)	461	29
	-	Worst(17.28/15.74)	248	53
Learning rate	1×10^{-4}	0.80/ 0.82	319	48
	1×10^{-5}	0.96/ 0.98	406	51
Momentum	0.9	-	-	-
Total network weight	475723	-	-	-

STM network collapse all processing into one dimensional CTC layer having 40 classes including a blank label which predict the output symbol. The performance was evaluated on various settings of proposed architecture as summarized in Table 2.

The performance comparison of said approach on handwritten, synthetic and scene text is detailed in Table 3. The offline and online handwritten Arabic is experimented by [12].

They presented their work in ICDAR 2009 handwriting competition. As presented in Table 3, they proposed hierarchical architecture. Later [20] used the same architecture by changing little bit in parameters like hidden memory blocks. Moreover, they experimented their work with MDLSTM networks. The details about their implementation can be found in their manuscript [20]. The presented approach on scene text using the hierarchical sub-sampling achieved benchmark accuracy in terms of Arabic scene text recognition.

a) Experimental Analysis

The experiments were conducted into manifold with various settings. The experimental settings were apparently outlined on the basis of architectural manipulation and parametric details. Following are the details of conducted experiments.

Table 3: Performance comparison of hierarchical subsampling on handwritten, synthetic and scene text Arabic script with pre-determined architecture.

Category	Epochs	Hidden Units	Output layer	Weights	sub-sample window	LSTM Dimension	Accuracy (%)
Online Handwritten Arabic [12]	85	20, 60, 180	CTC	423,926	[1],[2],[2]	1-DLSTM	95.70
Offline Handwritten Arabic [12]	91	4, 20, 100	CTC	550,334	[4,3],[4,2],[4,2]	2-DLSTM	95.70
Printed Arabic [20]	398	2, 10, 50	CTC	551,405	[4,3],[4,2],[4,2]	MDLSTM	98.25
Arabic Scene Text	406	20, 60, 100, 120	CTC	475,723	[4,3],[4,2],[4,2]	MDLSTM	96.81

1. The number of hidden layers were considered to investigate the performance of learning architecture.
2. The number of memory blocks at each layers using subsampled input.
3. The performance is explored by empirically selected learning rates.

As discussed earlier, that proposed network delegate the processing of MDL- STM network's learning to hidden layer units. The proposed method was evaluated on 3 and 5 hidden layer architecture.

At first, with three-layer architecture, each layer has 20. Then, by following same hidden layer architecture, each layer has 60 LSTM memory blocks and so on. Ultimately, with hidden layers size 3 and 5, the network was evaluated with each 20, 60, 100 and 120 LSTM memory blocks. Consequently, there are 8 experimental settings for each proposed architecture based on number of hidden layers as detailed in Table 4.

For the activation of the input and output unit used tanh whereas, function was used for gate's activation. The CTC layer has 38 output nodes for 37 input characters including one extra blank node. The 38 character input includes Arabic characters and numerals. All hidden layers in proposed architecture are fully connected to each other. The 3 hidden layer architecture was initially proposed where each layer was subsampled at first to 20 LSTM memory blocks. The performance was evaluated later on 40, 60, 80, 100.

Table 4: Details of performed experiments on 3 hidden layer architecture

Subsample size	Experiments	Hidden units/layer	Learning rate	Word Recognition Error(%)	Character Recognition Error (%)
6×6	Exp-1	20	1×10^{-4}	0.49	0.40
	Exp-2	60	1×10^{-4}	0.24	0.19
	Exp-3	100	1×10^{-4}	0.17	0.13
	Exp-4	120	1×10^{-4}	0.20	0.17
2×9	Exp-1	20	1×10^{-5}	0.55	0.51
	Exp-2	60	1×10^{-5}	0.33	0.23
	Exp-3	100	1×10^{-5}	0.09	0.06
	Exp-4	120	1×10^{-5}	0.24	0.16

Table 5: Details of performed experiments on 5 hidden layer architecture

Subsample size	Experiments	Hidden units/layer	Learning rate	Word Recognition Error(%)	Character Recognition Error (%)
6 × 6	Exp-1	20	1×10^{-4}	0.62	0.54
	Exp-2	60	1×10^{-4}	0.53	0.42
	Exp-3	100	1×10^{-4}	0.11	0.10
	Exp-4	120	1×10^{-4}	0.43	0.34
2 × 9	Exp-1	20	1×10^{-5}	0.59	0.48
	Exp-2	40	1×10^{-5}	0.31	0.24
	Exp-3	100	1×10^{-5}	0.19	0.12
	Exp-4	120	1×10^{-5}	0.22	0.14

The units defined in subsampled layers were also fully connected. The performance in hidden units were delegated backward to main hidden layers and the calculation of subsample layer was incorporated in the gradient descent of next hidden layer with learning rate 1 10−4 and then on 1 10−3 and momentum 0.9 which is selected after observing the trend from another cursive text analysis using MDLSTM. The training on each experiment was stopped after observing no significant improvement on performance for 30 epochs. Table 4 5

represent the details about number of epochs consumed for each experiment while the size of the hidden layer was 3 and 5. The learning rate and number of hidden sub-sampled layers on convolutional features are impacting the learning performance of training network. The output is presented in Figure 5. The recorded accuracy is 95.8% calculated by Levenshtein distance measure at character level as indicated in Table 6.

Figure 5: Observed scene text recognition output, the original input images were rescaled and converted into gray-scale. The output was mapped with ground truth. The green color symbols at output show insertions, whereas, deletions are presented in red color.

Table 6: Details of performed experiments on 5 hidden layer architecture

Error type	Test set Error
Deletions	43.75
Substitutions	41.91
Insertions	30.24

V. Conclusion and Discussion

The nature of Arabic script is extremely complex and cursive. To under- stand the Arabic word, there is a need to investigate the characters involved in predicting a word. The representation of characters is a considerable is- sue, because every character has four possibilities to occur in a word. The constraint of character's position make it difficult for any type of segmentation technique to correctly determine the characters by any specified technique. Therefore, there is always a need to look for implicit segmentation techniques that counter such complications associated to Arabic scripts. As Arabic script is a context-based language, hence context learning classifiers are suitable for learning purpose. The presented architecture for scene text analysis depicted good results. The obtained results exhibit that if there is a precise and relevant feature provided to learning network then it could pro- duce realistic results even on the intrinsic scripts. Experimental evaluation has also explained in detail which tells the learning trend and recognition accuracy at word and character level of Arabic scene text.

References Références Referencias

1. Saeeda Naz and Arif Iqbal Umar and Riaz Ahmad and Imran Siddiqi and Saad Bin Ahmed and Muhammad Imran Razzak and Faisal Shafait, "Urdu Nastaliq recognition using convolutional-recursive deep learning", in Neurocomputing (2017), vol: 243, pp: 80-87
2. Saeeda Naz and Arif Iqbal Umar and Riaz Ahmad and Saad Bin Ahmed and Syed Hamad Shirazi and Imran Siddiqi and Muhammad Imran Raz- zak, "Offline cursive Urdu-Nastaliq script recognition using multidimensional recurrent neural networks", in Neurocomputing in 2016, vol: 2016, pp. 228–241
3. Maroua Tounsi and Ikram Moalla and Adel M. Alimi and Frank Lebourgeois, "Arabic characters recognition in natural scenes using sparse coding for feature representations", ICDAR, ISBN: 978-1-4799-1805-8, pp: 1036-1040, url: "http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7321714
4. A. Graves and M. Liwicki and S. Fernandez and R. Bertolami and H. Bunke and J. Schmidhuber, "A Novel Connectionist System for Uncon- strained Handwriting Recognition", in IEEE Trans. Pattern Analysis and Machine Intelligence, vol: 31, (2009), pp. 855-868
5. Saad Bin Ahmed and Saeeda Naz and Salahuddin Swati and Muham- mad Imran Razzak, "Handwritten Urdu Character Recognition using 1- Dimensional BLSTM Classifier", in Neural Computing and Applications 2017.
6. Ryosuke Odate and Hideaki Goto, "Highly-accurate fast candidate reduction method for Japanese/ Chinese character recognition", in ICIP (2016), ISBN: 978-1-4673-9961-6, pp. 2886-2890
7. Saad Bin Ahmed and Saeeda Naz and Muhammad Imran Razzak and Rubiyah Yousaf, "Deep Learning based Isolated Arabic Scene Character Recognition", in 1st Workshop on Arabic Script Analysis and Recognition (2017), url: http://arxiv.org/abs/1704.06821
8. Saad Bin Ahmed, Saeeda Naz, Muhammad Imran Razzak, Rubiyah Yusof, "Evaluation of Handwritten Urdu Text by Integration of MNIST Dataset Learning Experience", in Neuro Processing Letters (NEPL).
9. Ahmed, Saad Bin and Naz, Saeeda and Razzak, Muhammad Imran and Yusof, Rubiyah and Breuel, Thomas M., "Balinese Character Recognition Using Bidirectional LSTM Classifier", in Advances in Machine Learning and Signal Processing, (2016), Springer International Publishing, pp:201– 211
10. Saeeda Naz and Saad Bin Ahmed and Riaz Ahmad and Muhammad Imran Razzak, "Zoning Features and 2DLSTM for Urdu Text-line Recognition", (2016), vol: 96, Procedia Computer Science, pp: 16-22
11. Saad Bin Ahmed and Saeeda Naz and Muhammad Imran Razzak and Shiekh Faisal Rashid and Muhammad Zeeshan Afzal and Thomas M. Breuel, "Evaluation of cursive and non-cursive scripts using recurrent neu- ral networks", in Neural Computing and Applications (2016), vol: 27, pp. 603-613.
12. Alex Graves, "Supervised Sequence Labelling with Recurrent Neural Networks", Springer Book, 2012, vol: 385, Studies in Computational Intelligence, ISBN:978-3-642-24796-5, pp: 1-131. url:http://dblp.uni-trier.de/https://doi.org/10.1007/978-3-642-24797-2",
13. Alex Graves and Santiago Fern´andez and Faustino J. Gomez and Ju¨rgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks", Ma- chine Learning, Proceedings of the Twenty-Third International Con- ference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25- 29, 2006, ACM International Conference Proceeding Series, vol: 148, ISBN = 1-59593-383-2, pp: 369-375, url: http://dblp.uni-trier.de/db/conf/icml/icml2006.htmlGravesFGS06",
14. Jonathan Fabrizio and Beatriz Marcotegui and Matthieu Cord, "Text segmentation in natural scenes using Toggle-Mapping", in IEEE ICIP 2009, ISBN: 978-1-4244-5654-3, pp. 2373-2376, url: http://dblp.uni-trier.de/db/conf/icip/icip2009.html FabrizioMC09
15. Z. Y. Zhang and L. W. Jin and K. Ding and X. Gao", "Character-SIFT: A Novel Feature for Offline

Handwritten Chinese Character Recognition", in ICDAR 2009, pp. 763-767.

16. Mohit Jain, Minesh Mathew and C. V. Jawahar," Unconstrained scene text and video text recognition for Arabic script", (2017), bibsource: http://dblp.uni-trier.de/db/journals/corr/corr1704. htmlAhmedNRY17, URL: http://arxiv.org/abs/1704.06821
17. Sonia Yousfi and Sid-Ahmed Berrani and Christophe Gar- cia," ALIF: A dataset for Arabic embedded text recognition in TV broadcast", in ICDAR 2015, IEEE Computer Society, ISBN: 978-1-4799-1805-8, pp.1221–1225, url: http://dblp.uni-trier.de/db/conf/icdar/icdar2015. html#YousfiBG15a",
18. Oussama Zayene and Jean Hennebert and Sameh Masmoudi Touj and Rolf Ingold and Najoua Essoukri Ben Amara, "A dataset for Arabic text detection, tracking and recognition in news videos- AcTiV", in ICDAR 2015, publisher:" IEEE Com- puter Society", ISBN : 978-1-4799-1805-8, pp : 996-1000, URL: http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7321714 http://www.computer.org/csdl/proceedings/icdar/2015/1805/00/index.html"
19. M. Ben Halima and H. Karray and A. M. Alimi, "Arabic Text Recognition in Video Sequences", in International Journal of Computational Linguistics Research. (2013), URL: http://arxiv.org/abs/1308.3243
20. Naz, Saeeda and Umar, Arif Iqbal and Ahmed, Riaz and Razzak, Muhammad Imran and Rashid, Sheikh Faisal and Shafait, Faisal, "Urdu Nasta'liq text recognition using implicit segmentation based on multi- dimensional long short term memory neural networks, SpringerPlus, 2016, vol: 5, ISSN: 2193-1801, url="https://doi.org/10.1186/s40064-016-3442- 4"