ADDIS ABABA UNIVERSITY

ADDIS ABABA INSTITUTE OF TECHNOLOGY

SCHOOL OF GRADUATE STUDIES

SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Correction of Distortion in Scene Text Recognition

By

Temesgen Mekonnen

April 2022

Addis Ababa, Ethiopia

ADDIS ABABA UNIVERSITY
ADDIS ABABA INSTITUTE OF TECHNOLOGY
SCHOOL OF GRADUATE STUDIES
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

Correction of Distortion in Scene Text Recognition

A thesis Submitted to the School of Graduate Studies of Addis Ababa University in partial fulfillment for the Requirements for the Degree of Masters of Science in Computer Engineering

By
Temesgen Mekonnen

Advisor
Dr. Surafel Lemma Abebe (Associate Professor)

April 2022

Addis Ababa, Ethiopia

Statement of Authorship

I the undersigned, declared this research work is my own original work. At the same time, I affirm that this research work has never been used for other academic degrees and all sources of materials and literary works referred for the research have been fully acknowledged.

_____

Temesgen Mekonnen

April 2022

Addis Ababa, Ethiopia

Certificate

This is to certify that this thesis work which is submitted by Temesgen Mekonnen is carried under my supervision and guidance and fulfilling the nature and standard required for the partial fulfilment of requirements of masters of science degree in Computer Engineering. The work embodied in this thesis has not been submitted elsewhere for degree.

_____

Dr. Surafel Lemma

School of Electrical and Computer Engineering
Addis Ababa Institute of Technology
Addis Ababa University

Date: _____

Addis Ababa University

Addis Ababa Institute of Technology

School of Electrical and Computer Engineering

The undersigned have examined the thesis titled:

Correction of Distortion in Scene Text Recognition

Presented by Temesgen Mekonnen Feleke, a candidate for the degree of Master of Science and hereby certify that it is worthy of acceptance.

Approved By Board of Examiners

Dr. Bisrat Derebssa Dean, School of Electrical and Computer Engineering	_____ Signature
Dr. Fitsum Asaminew External Examiner	_____ Signature
Dr. Bisrat Derebssa Internal Examiner	_____ Signature
Dr. Surafel Lemma Advisor	_____ Signature

April 2022

Addis Ababa, Ethiopia

Acknowledgement

Primarily, I would like to praise and thank God, the Almighty, who has granted countless blessing and opportunity to me, so that I have been finally able to accomplish the research.

First and foremost, I would like to express my sincere gratitude to my advisor Dr. Surafel Lemma, for his continuous support and supervision during my M.Sc. study and M.Sc. thesis research, for his patience, motivation, enthusiasm, and immense research knowledge. His guidance helped me in all the time of research and writing of this thesis. His timely and priceless constructive comments and suggestions throughout this research have contributed a lot for the success of this research.

In addition, I would like to acknowledge thesis examiners and fellow colleagues. I especially want to thank Dr. Fitsum A., Dr. Bisrat D., and Dr. Sreenivas N. for their encouragement, insightful comments, and alerting questions. I want to thank my fellow classmates at Addis Ababa University, Institute of Technology, School of Electrical and Computer Engineering. I also want to thank Mr. Turegne F. for your kind and friendly advices. Thanks also to anyone I've forgotten, who was instrumental in this thesis.

Last but not the least, I would like to thank my mother Kassaye K., for supporting me in every aspect throughout my life; my fiancé Silenat J., for your consistent support and love; my brothers and sisters; all of my friends and the rest of my family members. Finally, in memory of my beloved father Mekonnen F. I would like to express my sympathy, “Rest in Eternal Peace Dad”, Thank you very much for all what you have done as a best father!

Abstract

Scene texts which are found in scene images contain valuable meanings. Scene text recognition is the process of converting text regions on the image into machine readable and editable symbols. Naturally, scene texts can appear in regular or irregular layout. In scene texts irregular text is widely used. Scene text with an irregular layout is difficult to recognize because of different forms of distortions. Correcting these distortions without losing any desired information is one of the major challenges in computer vision. Different approaches are proposed to solve the problem of distortion in scene text recognition. Based on their proposed techniques, these approaches can be categorized into four categories: Character Level Strong Supervision, Rectification, Multi Direction Encoding and Attention based approaches. The state of the art is the attention-based approach which predicts characters from scene text image features by using encoder and decoder with attention methods. The performance of attention-based approaches, however, is not good mainly because they are unable to extract detail image features. The approach underperforms particularly with long scene text sequences due to their inconsistent and decreasing encoder output utilization during each decoding time step. Also, it faces the problem of attention mismatch for severely distorted texts.

To tackle the problem in attention-based encoder decoder approach, we proposed global attention based mechanism with Bi-LSTM decoder which can handle any type of text distortions implicitly. The proposed approach is trained with 6,000 regular and irregular scene text images randomly taken from publicly available SYN90K synthetic datasets. The dataset is widely used to train scene text recognizers. Preprocessing tasks which are image rescaling and noise removal are performed only for training purpose.

The proposed approach is evaluated using 4 class of regular scene text image datasets and 3 class of irregular scene text image datasets. The proposed approach outperforms the state-of-the-art approach by an average of 1.58% on regular scene text image datasets and by an average of 1.85% on irregular scene text image datasets. In addition, the incorporation of Bi-LSTM decoder in the proposed approach increases the recognition performance by an average of 5.24% for regular scene texts and by an average of 3.05% for irregular scene texts.

Keywords: Scene Text Recognition, Distortion, Irregular Scene Text Recognition, Computer Vision, Encoder, Decoder, Global Attention

Contents

List of Tables

List of Figures

Abbreviations

1D	One Dimensional
2D	Two Dimensional
AN	Attention Network
AON	Arbitrary Orientation Network
ASRN	Attention-based Sequence Recognition Network
ASTER	Attentional Scene Text Recognizer with Flexible Rectification
Bi-LSTM	Bi-directional Long Short-Term Memory
Char-Net	Character Aware Neural Network
CNN	Convolutional Neural Network
Conv	Convolution
CRNN	Convolutional Recurrent Neural Network
CTC	Connectionist Temporal Classification
CUTE (CT)	Curved Text
EOS	End of String
ESIR	End-to-End Scene-text recognition by Iterative Image-Rectification
FAN	Focusing Attention Network
FG	Filter Gate
FN	Focusing Network
ICDAR	International Conference on Document Analysis and Recognition
IC03	ICDAR 2003 Scene Text Image Dataset
IC13	ICDAR 2013 Scene Text Image Dataset
IC15	ICDAR 2015 Scene Text Image Dataset
IIIT5K	5000 Cropped Word Images from Natural Scene and Born-Digital Images

ImageNet	Large-scale Ontology of Annotated Images
LCD	Liquid Crystal Display
LSTM	Long and Short-Term Memory
MORAN	Multi Object Rectification Attention Network
MORN	Multi Object Rectification Network
OCR	Optical Character Recognition
PDAs	Personal Digital Assistances
RARE	Robust Scene Text Recognition with Automatic Rectification
ResNet	Residual Network
RGB	Red Green Blue
RNN	Recurrent Neural Network
R^2AM	Recursive CNN with Attention Modeling
SAR	Show-Attend-and-Read
SRN	Sequence Recognition Network
STAN	Sequential Transformation Attention-based Network
STAR-Net	A Spatial Attention Residue Network for Scene Text Recognition
STN	Spatial Transform Network
STR	Scene Text Recognition
SVT	Street View Text
SVT-P	Street View Text-Perspective
SWT	Stroke Width Transform
SYN90K	Synthetic Dataset based on 90 thousand Common English Words
SynSet	Synonym Set
tanh	Hyperbolic Tangent
TPS	Thin Plate Spline

Chapter 1

Introduction

Scene text is a textual content caught by a camera as an essential portion of scene images. Scene text constitutes the core semantics of natural scene images, which describes the name of organizations, brands, locations, nameplates, traffic warnings, road signs, etc. [6]. Scene text recognition is the process of converting text regions on the scene image into computer readable and editable symbols [10]. It is a wide area of research which is more challenging and extremely difficult since scene texts are often scattered in the scene image, and there is no prior information about their precise location. In addition, scene texts often have variety of font sizes, font types, backgrounds, colors, layouts and orientations [12].

Scene text recognition is gaining more interest in the research community. In recent years, the application of scene text recognition has expanded to computer vision tasks, such as autonomous driving, drone navigation, robot navigation, navigation for vision-impaired people, intelligent inspection, image searching and human-computer interaction [2].

In real-world scenarios, scene text often suffers from various types of distortions, such as perspective distortion or curvature distortion, which increases the difficulty of recognition [1]. Nowadays, regular scene text recognition methods have achieved notable success [3]. However, most current scene text recognition methods remain too unstable to handle multiple disturbances from the environment [3]. The various shapes and distorted patterns of irregular scene texts are the main challenges in recognition, such as perspective and curved text [3], [23]. Despite a lot of promising research work on regular scene text recognition, recognizing irregular scene texts from natural scene images is still a challenging task and its performance is far from being satisfactory [21].

1.1 Problem Statement

The accuracy of scene text recognition is affected by many factors. Among these factors distortion is one of the major challenges. Distortion is irregular appearance of scene texts on natural

scene images and it can be perspective or curvature. Distorted scene texts are common in natural scene images and recognizing them is difficult [1]. Therefore, in order to recognize them correctly, the distortion should be corrected and handled properly. Different methods are proposed to solve the problem of distortion. Based on their approaches these methods can be grouped into four categories: character level strong supervision [15], [34], [47], rectification [1], [2], [3], [22], [23], multi direction encoding [21] and attention [17], [19], [20] based approaches.

In scene text recognition of irregular scene text with distortions, the state of the art is attention-based encoder-decoder approach that learns the mapping between input scene text images and output character sequences in a purely data-driven way from scene text image features [19], [20], [27]. However, existing attention-based methods underperform mainly because (1) they are unable to extract detail features; and (2) they have poor encoder output utilization and attention mismatch due to misalignment between feature areas and targets for severely distorted and long scene texts.

Therefore, the purpose of this research is to develop an approach which can address these problems and achieve better performance. In this research, based on attention-based encoder decoder framework, we propose global attention mechanism which addresses the problems to handle any type of scene text distortions implicitly without explicit scene text image transformation or rectification. In addition, we incorporate Bi-LSTM decoder instead of one directional LSTM decoder for boosting scene text recognition performance. This research work aims to answer the following research questions.

- RQ1 [Global Attention Mechanism]

Can we improve the overall scene text recognition performance by using global attention mechanism in regular and irregular scene text recognition?

- RQ2 [Bi-LSTM Decoder]

What is the effect of using Bi-LSTM decoder instead of using one directional LSTM decoder on the performance of regular and irregular scene text recognition?

1.2 Objective

1.2.1 General Objective

The main objective of this research is to address the problem of scene text distortions. In particular we aim to implement a scene text recognition system by using global attention mechanism to solve the problem.

1.2.2 Specific Objectives

- Investigate various prior approaches proposed to address the problem of distortion in scene text recognition.
- Propose and develop global attention based encoder decoder approach to handle irregular scene texts with various types of distortions.
- Demonstrate an explicit effect of Bi-Lstm decoder on performance of regular and irregular scene text recognition.
- Selecting and preprocessing scene text image datasets.
- Evaluate the performance of the proposed approach and compare it with the state-of-the-art approach.

1.3 Scope

The focus of this research is to develop a scene text recognition system that can handle regular scene texts and irregular scene texts with perspective or curvature distortions. Vertical scene texts and scene texts with calligraphic fonts are not considered in this research.

1.4 Contributions

This research extends the field of scene text recognition from natural scene images with the following contributions:

- Develop the first scene text recognition system based on global attention mechanism with encoder and Bi-LSTM decoder which can handle scene text distortions implicitly.
- Reproduce the baseline work called SAR [19] which can be used for comparison in future related researches on correction of distortion in scene text recognition.
- This research plays a great role in understanding the challenge of correction of distortion in scene text recognition.
- This research shows a new research direction towards further development of language translation systems and text to voice conversion systems which can assist tourists, self-driving cars and blind peoples.

1.5 Research Methodology

In this research work, the first task was conducting a literature review. Throughout the research different literatures have been reviewed based on state-of-the-art approach to identify gaps in previous works and formulate the problem statement. Once a problem is formulated, an approach which can be a solution for handling the stated problems is proposed and implemented. After implementing the proposed approach, scene text image datasets are collected and preprocessed accordingly to train the proposed approach. Finally, the proposed approach is evaluated using evaluation benchmarks and compared with the state-of-the-art approach.

1.6 Thesis Organization

The rest of this document is organized as follows. Theoretical backgrounds which include definitions, text types, scene text recognition tasks, challenges of scene text recognition and applications of scene text recognition are discussed in Chapter Two. Chapter Three presents summary of previous works proposed to handle the problem of distortion in scene text recognition. Our proposed approach is explained in Chapter Four. Experimental scenarios, evaluation of the proposed approach and results and discussions are presented in Chapter Five. Finally, Chapter Six summarizes the research and identifies gaps which can be addressed in the future.

Chapter 2

Theoretical Background

A text is anything that can be read and conveys a set of meanings to the person who examines it (Yuri Lotman, 1975). Text used as a carrier of high-level semantics [10]. It carries rich and valuable semantics to convey useful information for the reader. The reader can be a human or a machine.

2.1 Text Detection and Recognition

Text detection is the process of predicting the presence of text and localizing each available instance of text in a given text material [10], [12], [34]. It can be performed at character level, word level or sentence level. While text recognition is the process of converting text regions on the given text material into machine readable and editable symbols [10], [15], [40]. The text material can then be a document. In this work our focus is on scene text recognition rather than text detection.

There are two general approaches for text recognition in any text material which are bottom-up approach and top-down approach [17]. In bottom-up approach characters are first detected and recognized one by one using techniques, such as sliding windows [40], connected components [45] or Hough transforms [41]. The recognized characters are then integrating into the output text. In top-down approach the entire text is directly recognized from the original text material without detecting the characters one by one.

Text can be classified as document text or scene text. Document text is a text available on the given document as illustrated in Figure 2.1 (a). While scene text is a text available on the given scene image as illustrated in Figure 2.1 (b). Document text has clean background, regular and uniform font type and font size, plain layout and monotone (single) color. Hence, they can be simply recognized by traditional optical character recognition (OCR) techniques [1]. However, scene text has complex and cluttered background, distinct font type and size, complex layouts, various orientations, perspective or curvature distortions and different colors [10]. In addition,

there are a number of interference factors like noise, blur, non-uniform illumination, low resolution and partial occlusion. Hence, it is a bit challenging to detect and recognize them [20]. Despite several decades of research on document text recognition using OCR, recognizing texts from natural scene images is still a challenging task and its performance is far from being satisfactory [17], [19], [33].

Figure 2.1: Document Text (a) with Clean Background; and, Scene Texts (b) with Complex Background

2.1.1 Document Text Recognition

Document text recognition is the process of converting scanned images of hand written, type written or printed documents into machine readable format [48]. Document text can be scanned, typed or printed paper documents like newspaper, report, letter, article, memorandum, sales invoices, and any other form of documents.

Document texts with professional scan quality can be recognized by OCR technology up to nearly 100% recognition accuracy [17]. Despite OCRs' remarkable success in handling document texts, when we apply it for recognizing natural scene texts, its recognition performance on average falls below 40% [6], [40]. This happens since inputs of most OCR engines are well-segmented and binarized texts which have been easily distinguished from the background. However, the segmentation of scene text is far harder and complex for natural scene text images leading to poor

recognition results [6]. Therefore, we need another way of thinking to tackle the problem of scene text recognition.

2.1.2 Scene Text Recognition

Scene texts are naturally existing texts as part of images and videos [6], [50]. They are textual contents caught by a camera as an essential portion of images and videos, which shows the existence of valuable semantics as an information. Some of their exemplary contents are the names of a restaurants, commercial advertisements in stadiums and LCD display screens, names of supermarkets, names of governmental and non-governmental institutions, product container leveling names, nameplates, traffic warnings and road signs [6].

Scene text recognition is the process of converting texts on natural scene images into machine readable and editable symbols [4], [6]. Compared to document texts, recognizing scene texts has many additional challenges due to the following unique characteristics [6], [17].

- In scene images, many different objects like building structures, vehicles, leaves of trees, humans, animals, or parts of them may have similar shape and appearance to the text. Therefore, scene texts found within objects with cluttered background can be mis-recognized. For example, corner of a building can easily be recognized as character I or wheels of a vehicle can be recognized as character O. This makes discriminate text parts from non-text parts more difficult.
- Most of the time scene texts have decorative and stylish fonts usually to attract the attention of viewers. Unlike humans, to identifying and recognizing such ambiguous and decorated characters is challenging mainly because training is done on undecorated normal font types. In addition of this, scene texts can have any orientations and come with challenges like low contrast, blur and low resolution.
- Scene texts have no clean and standard layout. Specific scene text regions and amount of text within it vary from one scene image to the other. This makes text region localization and word or line segmentation harder than usual OCR document images.

- There is also variability of scene text size, type, color, orientation and alignment even they are part of the same word. Sometimes part of the text may be occluded by objects or camera standing positions.

Before the advent of deep learning techniques, scene text recognition is performed with lexicon-based methods which retrieves the nearest word from the given lexicons by calculating the word distance probability distributions [1], [3]. However, after the introduction of deep learning techniques, there is no mandatory need of lexicons.

Scene text recognition tasks can be further classified as regular scene text recognition and irregular scene text recognition.

2.1.2.1 Regular Scene Text Recognition

Regular scene texts are scene texts which are frontal and parallel [19]. They have almost uniform and horizontal arrangements. There is no text irregularities and any form of distortions with them. Therefore, recognizing such type of scene texts is known as regular scene text recognition. Most of the existing works focused on the recognition of regular scene texts [7].

2.1.2.2 Irregular Scene Text Recognition

Irregular scene texts are scene texts with many forms of text irregularities and perspective or curvature distortions. In scene images irregular scene texts are widely used for different purposes. Recognizing such type of scene texts is known as irregular scene text recognition. However, scene text with an irregular layout is difficult to recognize because of its distorted patterns [3].

2.2 Application of Scene Text Recognition

Scene text recognition is a hot research area in computer vision community [2]. It plays an important role in several applications such as autonomous self-driving cars, drone navigation, robot navigation, navigation for vision-impaired people, intelligent inspection, product search, vehicle license recognition, instant translation, industry automation, image searching and human-computer interactions [3], [33].

In computer vision community where fast development of camera-based applications are readily available on mobile devices like smart phones, tablets and personal digital assistances (PDAs), understanding scene text is much more important than ever. Applications that are available to answer questions such as, 'what is the meaning of scene texts available on this image?' are becoming increasingly popular [6]. However, scene text recognition is not an easy task. It is difficult to develop a general approach to recognize scene texts from scene images [1], [20].

2.3 Common Challenges of Scene Text Recognition

There are challenges which are common in the field of scene text recognition as a whole. Below, we list four challenges. Figure 2.2. shows examples of these common challenges. These challenges are still open for researchers in the field.

- **Calligraphic Fonts: -** Calligraphic fonts are more artistic fonts than the average fonts. They are often used for brands, like the brand name of Coca Cola. Recognizing calligraphic fonts correctly is still a big challenge.
- **Vertical Scene Texts: -** Vertical scene texts are texts with vertical orientations from top to bottom or vice versa. Still recognizing vertical scene texts is a big challenge in the field of scene text recognition.
- **Occlusion: -** Heavy occlusion can hide parts of the scene text which are available in the scene image. Texts available on this type of scene images are difficult to recognize.
- **Low Resolution: -** Low resolution means inadequate number of pixels per each patch of the scene image. Therefore, in order to recognize texts on this type of scene images prior resolution boosting mechanism is needed. Hence, recognizing them directly is difficult.

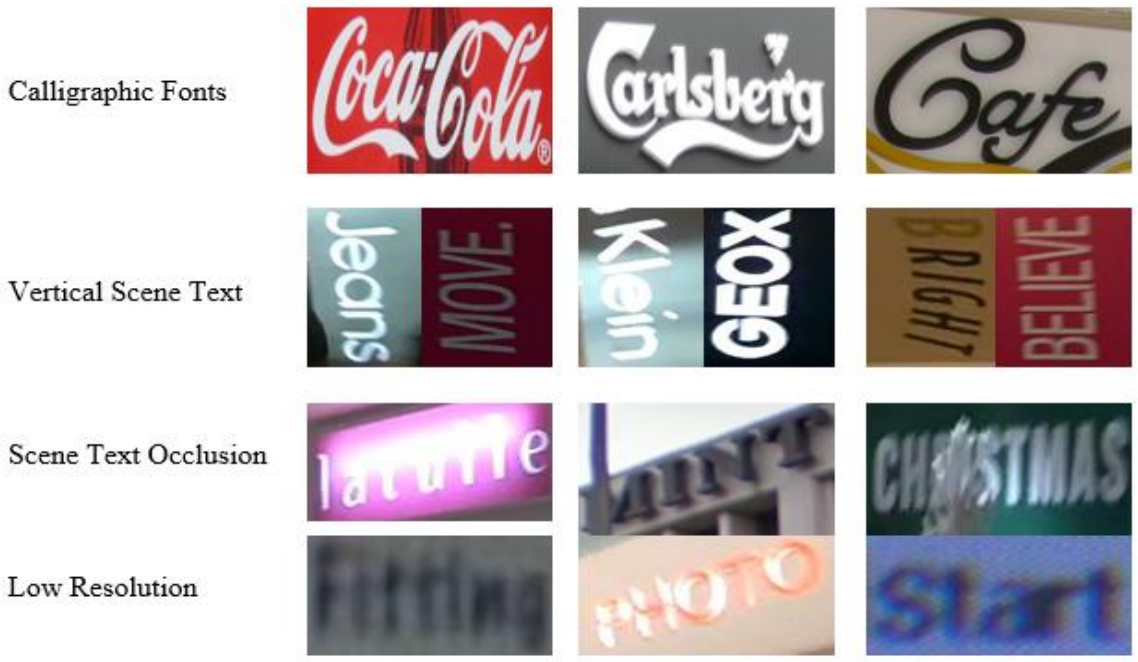


Figure 2.2: Common Challenges in Scene Text Recognition

In this research we don't use scene text images with calligraphic fonts, vertical orientations and occlusion for training purpose. Scene texts which suffers from a problem of low resolution becomes blurry and appears both in training and testing datasets. For training datasets we try to mitigate this blurry image problem by using a median filter to preserve better text line edges before passing them to a feature extractor network. However, since we use the common ICDAR standard 7 benchmarks for evaluation purpose, there are some such scene text instances in some evaluation dataset categories which are applied for testing the state-of-the-art and our proposed approach uniformly. Therefore, they can't affect the comparison validity and result interpretation as a whole.

Chapter 3

Literature Review

In this section previous research works for correction of distortion in scene text recognition are presented. Several methods are proposed to solve the problem of distortion in scene text recognition. Based on the proposed techniques, these approaches can be categorized into four: Character Level Strong Supervision, Rectification, Multi Direction Encoding and Attention based approaches. Character level strong supervision based approach is a traditional approach based on character level annotations [34], [47]. Rectification based approach involves different image transformation techniques [1], [2], [23]. Multi direction encoding based approach extracts four direction features and use character placement clues [21]. Attention based approach predicts characters from image feature maps by using encoder and decoder with attention methods [17], [19], [20]. The following sections explain these approaches in detail.

3.1 Character Level Strong Supervision Based Approach

Character level strong supervision based approach is a traditional approach based on binarization-based methods of character segmentation of scene text images, optical character recognition (OCR) and character level annotations.

Jaderberg et al. [34] proposed a traditional method of scene text recognition by using optical character recognition (OCR) and text spotting methods. Text spotting method identifies text regions on the scene image by suppressing non text regions of the scene image. Since this method proposed before the era of deep neural networks, it neither uses convolutional neural network (CNN) nor recurrent neural network (RNN). Instead, it uses stroke width transform (SWT) algorithms to find character candidates that are found on the scene image, filters the scene image, and applies text line aggregation to make accurate word detection. It finally applies binarization-based character segmentation and sliding-window-based OCR methods on the detected word region to recognize character sequences. Although it uses manually labeled characters as its lexicons, character labeling is tedious and consumes more resources and time. In addition, text

spotting methods like SWT performs well for document texts, since most of the time, document texts have uniform and similar character width and horizontal layout but scene text characters have various orientations, styles and non-uniform widths. Therefore, this method under performs for scene text recognition tasks.

Liu et al. [47] proposed a character aware neural network (Char-Net) for distorted scene text recognition. It is composed of character-level scene text image segmentation and connectionist temporal classification (CTC) based recognizer. Unlike rectification-based methods which rectifies the entire distorted scene images, it detects, rectifies and recognizes individual characters after segmentation is performed. It is analogues to document text recognition since it uses optical character recognition (OCR) and treats each individual characters as a sequential object. Finally, it applies dynamic programming technique to assemble character sequences to get meaningful words as a final recognized text output, in a bottom-up fashion. Dynamic programming techniques add another level of complication for the task of scene text recognition.

In general, character level strong supervision based approaches require the binarized segmentation of each character on the scene image, which is very challenging because of the complicated background clutter and distinct inline distances between consecutive characters. Moreover, if the scene text layout is irregular, the character segmentation becomes challenging and in some cases it becomes impossible. In addition to segmentation, character level annotations consume additional resources and time.

Method	IC03	IC13	IC15	IIIT5K	SVT	SVT-P
Bissacco et al. (Bissacco et al. 2013)	-	87.6	-	-	78.0	-
*Jaderberg et al. (Jaderberg et al. 2016)	*93.1	*90.8	-	-	*80.7	-
Jaderberg et al. (Jaderberg et al. 2015a)	89.6	81.8	-	-	71.7	-
Lee et al. (Lee and Osindero 2016)	88.7	90.0	-	78.4	80.7	-
(Shi, Bai, and Yao 2016) (CRNN)	89.4	86.7	-	78.2	80.8	66.8
(Shi et al. 2016) (RARE)	90.1	88.6	-	81.9	81.9	71.8
(Liu et al. 2016) (STAR-Net)	89.9	89.1	-	83.3	83.6	73.5
Char-Net	91.5	90.8	60.0	83.6	84.4	73.5
*Char-Net	93.3	93.7	-	-	87.6	-

Table 3.1: Scene Text Recognition Accuracies (%) of Char-Net on Public Benchmarks [47]. "*" Denotes Constrained Output to 90k Word Lexicons

3.2 Rectification Based Approach

Rectification based approaches attempt to transform irregular scene images into regular ones with explicit rectification mechanisms then recognize them using any kind of regular text recognizers.

Shi et al. [23] and Liu et al. [22] proposed a baseline approach based on deep neural network, which consists of a spatial transformer network (STN) for rectification and a sequence recognition network (SRN) for recognition as illustrated in Figure 3.1. In this approach, scene image is rectified first via a predicted Thin-Plate-Spline (TPS) transformation [51] technique into a more regular and readable scene text image for the following SRN, which recognizes scene texts through a sequence recognition method. For sequence recognition network, this approach uses connectionist temporal

classification (CTC) method [45] to recognize and classify characters to a number of unique character classes.

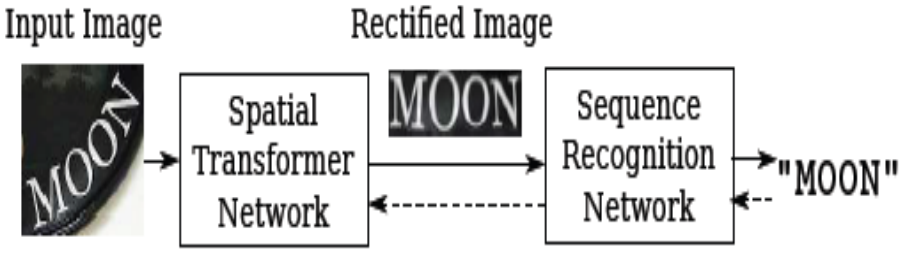


Figure 3.1: Rectification Approach with Spatial Transformer Network (STN) and Sequence Recognition Network (SRN) [23]

Connectionist Temporal Classification [45] (CTC) provides a differentiable loss function that is insensitive to horizontal character placement and spacing, enabling end-to-end trainable sequence recognition. It treats each character as a separate sequential object. Therefore, despite its effectiveness, it does not have a mechanism to model the dependencies among its output characters. Hence, Liu et al. [22] and Shi et al. [23] relies on an external language model, such as a lexicon, to incorporate language priors into their recognition network. Therefore, Shi et al. [1] modified this baseline approach by introducing attention-based encoder decoder as a text recognition model to replace CTC and achieves a greater scene text recognition performance than Liu et al. [22] and Shi et al. [23] (see Table 3.2).

Shi et al. [1] uses convolutional-recurrent neural network (CRNN) model for encoder and attentional sequence to sequence LSTM model for decoder. The encoder extracts a sequential feature representation from the rectified input scene image. The decoder recurrently generates a sequence of characters conditioned on the sequential feature representation, by decoding the relevant contents it attends to, at each time step. Since the output of a sequence-to-sequence model is generated by a recurrent neural network, it captures the output character dependencies, thus it incorporates language modeling into its recognition network. Hence the proposed approach by Shi et al. [1] does not need external language models such as a lexicon. If the fiducial points (rectification control points) are regressed incorrectly, Shi et al's [1] approach may incompletely

crop a text region from the scene text image. This leads to information loss of scene text, and hence, it results in poor performance in overall scene text recognition accuracy.

Luo et al. [3] proposed a multi object rectified attention network (MORAN) for scene text recognition. MORAN consists of a multi object rectification network (MORN) and attention-based sequence recognition network (ASRN) as shown in Figure 3.2. MORN is designed for rectifying scene images that contain irregular text. It rectifies every character object in the scene image independently. ASRN focuses on each target characters and produces sequential output character predictions. Moreover, to obtain accurate alignments between feature areas and targets, and also to improve sensitivity of the attention-based sequence recognition network, it proposes a fractional pickup method to train attention-based decoder in the ASRN. As we see from Table 3.2, Luo et al. [3] outperforms Shi et al. [1] only on regular scene text datasets of IC03 and IC13. Therefore, Shi et al. [1] is better particularly for irregular scene text recognition tasks. In addition, training process of MORAN is difficult, unstable and can't be done end-to-end in one step. Hence, it proposes a three-step curriculum learning strategy which makes the model complex and inefficient without significant performance improvement.

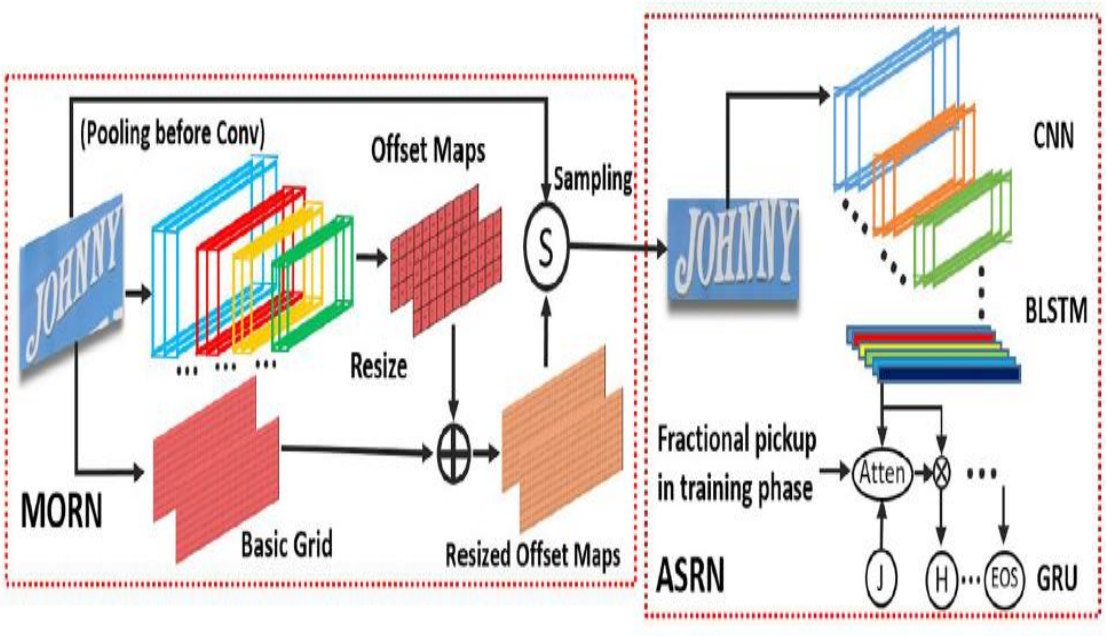


Figure 3.2: Overall Structure of MORAN [3]

Zhan et al. [2] proposed an iterative rectification-based approach (ESIR) which is trainable end-to-end and iteratively removes perspective and curvature distortions as illustrated in Figure 3.3. It introduces an innovative rectification network with a novel line-fitting transformation technique to estimate the appearance of text lines in scene images. In addition, it develops iterative rectification pipeline where scene text distortions are corrected iteratively towards an accurate and frontal-parallel view. It outperforms both Shi et al. [1] and Luo et al. [3] in all irregular scene text datasets of IC15, CT80 and SVT-P (see Table 3.2). However, Zhan et al. [2] rectifies irregular scene text images several times to come up with an acceptable final result and it takes enormous amount of time and space. In a recent work, Lin et al. [5] proposed a sequential transformation attention-based network (STAN), which comprises a sequential transformation-based rectification network and an attention-based recognition network. The sequential transformation network which is based on affine transformation [52], rectifies irregular scene text image by decomposing the task into a series of patch-wise basic transformations, followed by a grid projection submodule to smooth the junction between neighboring patches. STAN of Lin et al. [5] improves the rectification flexibility by applying spatial transformation network (STN) on divided and decomposed scene image patches rather than on the whole scene image. It achieves complex transformation by applying several simple but efficient transformations. Compared with Luo et al. [3], it is stable, more flexible and can be trained in an end-to-end manner. Compared with Zhan et al. [2], it doesn't require iterative rectification, and can achieve a good result in one rectification (see Table 3.2). However, it is unable to handle severely curved text. Therefore, it underperforms than Zhan et al. [2] for recognition of curved text dataset of CT80 as we can see from Table 3.2. Moreover, affine transformation is limited to scaling, rotation and translation.

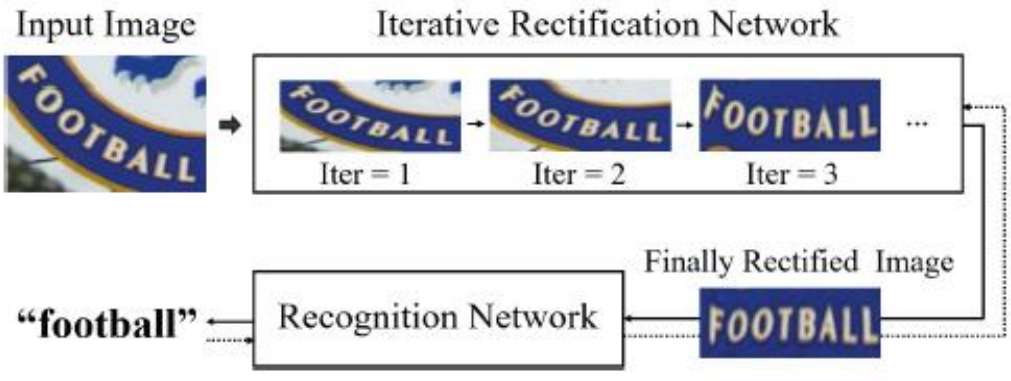


Figure 3.3: Overall Structure of ESIR [2]

In general, rectification based approaches are unable to rectify text when the input scene image is severely curved or distorted. This difficulty comes from the transformation techniques used for rectification tasks. They used transformation techniques like affine transformation, Euclidean transformation, thin-plate-spline (TPS) transformation and different line fitting transformations which all are inflexible and have geometric constraints. Moreover, they are not optimized without human labeled geometric ground truth.

Method	IC03	IC13	IC15	IIIT5K	SVT	SVT-P	CT80
Lee et al. [20]	88.7	90.0	-	-	-	-	-
Cheng et al. [17]	91.5	89.4	66.2	83.7	82.2	71.5	63.9
Shi et al. [23] (RARE)	90.1	88.6	-	81.9	81.9	71.8	59.2
Shi et al. [1] (ASTER)	94.5	91.8	76.1	93.4	93.6	78.5	79.5
Liu et al. [22] (STAR-NET)	89.9	89.1	-	83.3	83.6	73.5	-
Jaderberg et al. [40]	93.1	90.8	-	-	80.7	-	-
Cheng et al. [21]	91.5	-	68.2	87.0	82.8	73.0	76.8
Luo et al. [3] (MORAN)	95.0	92.4	68.8	91.2	88.3	76.1	77.4
Zhan et al. [2] (ESIR)	-	91.3	76.9	93.3	90.2	79.6	83.3
Lin et al. [5] (STAN)	95.1	92.8	76.7	94.1	90.6	82.2	82.3

Table 3.2: Scene Text Recognition Accuracies (%) of ESIR, MORAN and other Rectification based related works on Public Benchmarks

3.3 Multi Direction Encoding Based Approach

Multi direction encoding based approach extracts four direction features and use character placement clues.

Cheng et al. [21] proposed arbitrary orientation network (AON) to directly capture the deep features of irregular scene texts, which are combined into an attention-based decoder to generate character sequences. The whole network can be trained end-to-end by using only scene images and word-level annotations. It was inspired by the fact that, in previous methods any scene text available on scene images is treated in the same direction, by default from left to right but this might not be true in the wild. Therefore, it suggests that, the visual representation of an arbitrarily oriented character in a two-dimensional scene text image can be described in four directions by using four-character placement clues, which are left to right, right to left, top to bottom and bottom to top as illustrated in Figure 3.4. It also designs a filter gate (FG) for fusing redundant, four direction features to generate the integrated feature sequence by neglecting irrelevant features. It can recognize both regular and irregular scene texts from scene images. However, the model is extremely complex and uses sophisticated design.

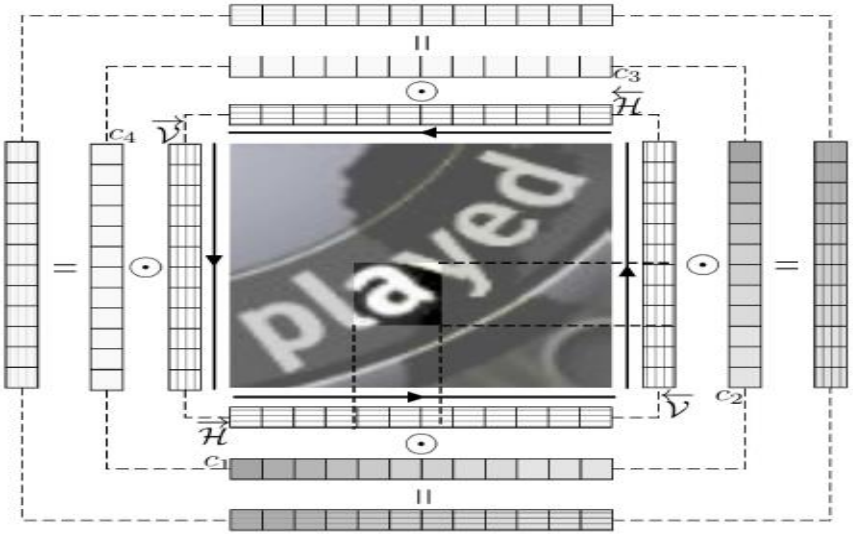


Figure 3.4: Four-Directional Character Visual Representation of AON [21]

3.4 Attention Based Approach

Attention based approaches predict and generate character sequences from extracted feature maps of scene text images by using a pair of encoder and attentional decoder. Attention-based mechanisms can allow the model to focus on the most important patterns of incoming features by ignoring the least important ones, as well as potentially adding a degree of interpretability [46].

Lee et al. [20] proposed recursive convolutional neural networks with attention modeling (R^2AM) for lexicon-free optical character recognition in natural scene images. This model implicitly learns sequential dynamics of character level language models that are naturally present in word strings. Moreover, it embodied in a recurrent neural network which avoids the need to use N-gram language models and lexicons. It uses soft attention mechanism which allows the model to selectively exploit scene image features in a coordinated way and can be trained end-to-end within a standard backpropagation framework. Since it uses one-dimensional attention modules, it is effective when dealing with sequential data which have frontal and parallel layouts. For regular scene text recognition tasks, recognizing characters in scene images can be considered as a task of recognizing horizontally sequential data. Therefore, it achieves good performance particularly for regular scene text recognition tasks. However, despite its success, it has some problems. For instance, the use of recursive CNNs for scene image feature extraction requires a relatively large memory, computation cost and time. In addition, its one-dimensional attention module is inefficient for irregular scene text recognition tasks. Lee et al. [20] is not evaluated on irregular scene text benchmarks.

Cheng et al. [17] proposed an approach which is called the focusing attention network (FAN) that employs a focusing attention mechanism to automatically adjusts the drifted attention. It was motivated by the fact that, attention regions in attention network deviate in some degrees from the correct target character regions on the scene images. Different from the existing methods, it adopts a ResNet-based convolutional neural network for the first time to extract deep feature representations of scene text images. As illustrated in Figure 3.5, FAN consists of two major components which are an attention network (AN) that is responsible for recognizing character sequences as in the existing methods, and a focusing network (FN) that is responsible for automatically adjusting the attention by evaluating whether or not AN pays attention properly on the target areas in the scene image.

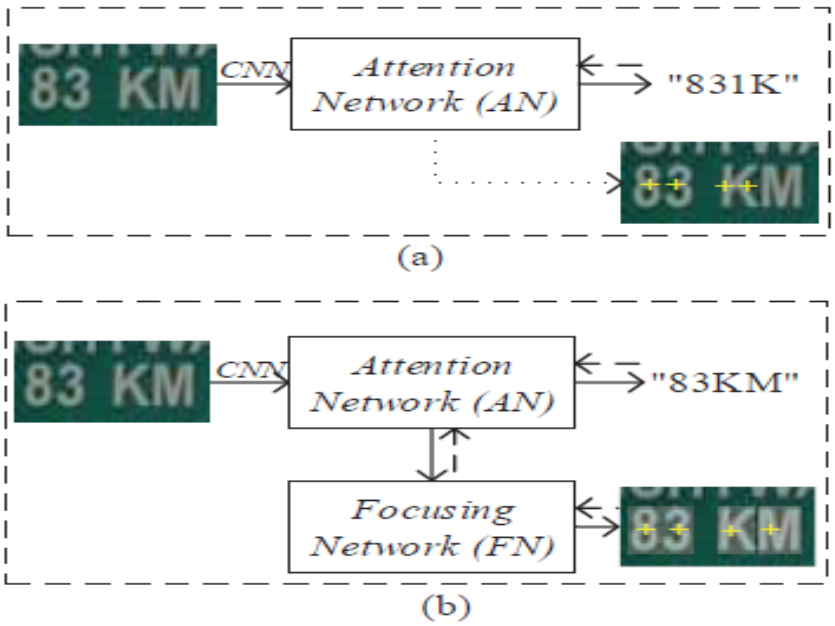


Figure 3.5: Overall Structure of FAN [17]

Generally, focusing network (FN) is used to evaluate whether the context vectors are reasonable or not, and provides feedback to help attention network (AN) to generate more reasonable context vectors so that the AN can pay attention properly on the correct and right regions of target characters in the processed scene images. The focusing mechanism of FN works in two major steps: 1) computing the attention center of each predicted labels; and, 2) focusing attention on target regions by generating the probability distributions on the attention regions. However, Cheng et al. [17] needs to be trained with character level bounding box annotations. Hence, it cannot recognize word strings directly with word level annotations. Therefore, its character level annotation adds another level of complexity and computation overhead.

Inspired by the success of Xu et al. [46], in the field of image captioning, Li et al. [19] adopts a two-dimensional (2D) local attention-based encoder decoder method which is known as show-attend-and-read (SAR). Xu et al. [46] designed for image captioning but Li et al. [19] used for scene text recognition. Li et al. [19] extends the one-dimensional (1D) attention-based encoder-decoder framework used in Lee et al. [20] and Liu et al. [22] to two-dimensional (2D) attention-based encoder-decoder framework that used to handle the complicated spatial layout of irregular scene texts. This is the first work that uses 2D attention mechanisms for irregular scene text recognition task.

Method	IC03	IC13	IC15	IIIT5K	SVT	SVT-P	CT80
(Jaderberg et al. 2015a)	-	90.8	-	-	-	-	42.7
(Lee et al. [20]) (R^2AM)	88.7	90.0	-	78.4	80.7	-	-
(Wang and Hu 2017)	-	-	-	80.8	81.5	-	-
(Shi et al. [23]) (RARE)	90.1	88.6	-	81.9	81.9	71.8	59.2
(Liu et al. 2016)	-	89.1	-	83.3	83.6	73.5	-
(Shi, Bai, and Yao 2017)	-	89.6	-	81.2	82.7	66.8	54.9
(Yang et al. 2017)*	-	-	-	-	-	75.8	69.3
(Cheng et al. [17])* (FAN)	94.2	93.3	85.5	87.4	85.9	71.5	63.9
(Liu et al. [47])* (Char-Net)	91.5	90.8	60.0	83.6	84.4	73.5	-
(Liu, Chen, and Wong 2018)*	-	91.1	74.2	92.0	85.5	78.9	-
(Bai et al. 2018)*	-	94.4	73.9	88.3	87.5	-	-
(Cheng et al. [21]) (AON)	91.5	-	68.2	87.0	82.8	73.0	76.8
(Shi et al. [1]) (ASTER)	94.5	91.8	76.1	93.4	93.6	78.5	79.5
(Li et al. [19]) (SAR)	-	94.0	78.8	95.0	91.2	86.4	89.6

Table 3.3: Scene Text Recognition Accuracies (%) of AON, FAN, SAR and other Attention based related works on Public Benchmarks [19] and [21]. The approaches marked with "*" are trained with both word-level and character-level annotations

As illustrated in Figure 3.6, it is simple, flexible and robust in handling different scene text layouts and orientations. However, it has three main problems. First it is unable to represent detail image features. Second, since it has poor encoder output utilization during each decoding time step, it

underperforms with long scene text sequences. Similar to Cheng et al. [17] approach faces the problem of attention mismatch (misalignment between feature areas and targets for distorted scene texts). The mismatch is due to the context vector which is calculated from the preliminary feature map representations.

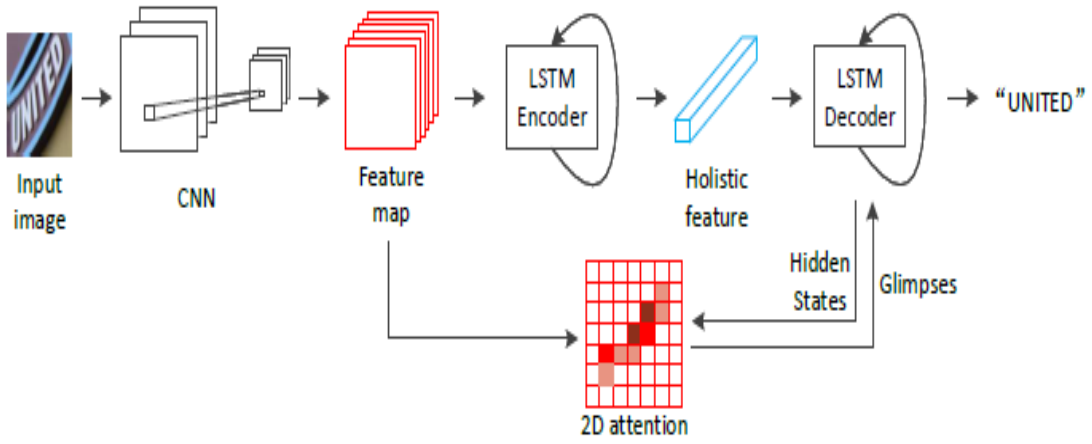


Figure 3.6: Overall Structure of SAR [19]

From all the literatures reviewed, we conclude that, character level strong supervision based approaches require binarization based segmentation of characters on the scene image, which is very challenging because of the complicated background clutter and distinct inline distances between consecutive characters. Moreover, if the scene text layout is irregular, the character segmentation becomes challenging and in some cases it becomes impossible. In addition to segmentation, character level annotations are tedious and consume additional resources and time. To solve these problems researchers come up with a new rectification based approach. However, due to the transformation techniques used for rectification task, this approach is unable to rectify text when the input scene image is severely curved or distorted. It also used transformation techniques like affine transformation, Euclidean transformation, thin-plate-spline transformation and different line fitting transformations which all are inflexible and have geometric constraints. Moreover, they are not optimized without human labeled geometric ground truth. Next to this, to extract 4 direction features of scene images to achive better recognition performance multi direction encoding based approach is proposed. But, it is extremely complex and uses sophisticated design without significant performance improvement. Finally attention based approach come to

reality after the introduction of attention mechanisms in artificial neural network. However, most works in this category unable to extract detail image features and have poor encoder output utilization during each decoding time step due to their local attention mechanism. Hence they underperform with long scene text sequences. Moreover, due to naturally existing attention drifts, this approach faces the problem of attention mismatch particularly for severely distorted scene texts. Therefore, based on existing attention based encoder decoder framework we propose global attention based approach that can handle these problems in a better way.

Most of the works which are mentioned in the literature review are trained by using the 9-million SYN90K training dataset. However, in this research we select and use 6,000 scene images from SYN90K for training purpose of our own work and the reproduced baseline work.

Chapter 4

Proposed Approach

In this section we describe the architecture of our proposed approach. The proposed approach takes scene images as an input and produces a varying length sequence of characters. General architecture of the proposed approach and other similar works under attention based approach is depicted in Figure 4.1. This architecture consists of four core modules which are:

- Feature Extraction Network
- Encoding
- Attention Mechanism
- Decoding

The state-of-the-art Li et al. [19] uses ResNet-31 for feature extraction network. In this work we use ResNet-32 as a feature extraction network for both the proposed approach and the reproduced work Li et al. [19]. As shown in Figure 3.6 Li et al. [19] utilizes the final holistic feature output of the Lstm encoder only once to initialize the Lstm decoder. But in this work we utilize both the holistic feature output of the Lstm encoder which used to initialize the Lstm decoder and combination of all time step hidden state outputs of the Lstm encoder for calculating every time steps' context vector to attend and monitor every time step decoders' output as shown in Figure 4.6. As an attention mechanism Li et al. [19] uses 2D local attention mechanism which calculates every time step context vector from the previous time step decoders' hidden state output and the local feature representation of scene images from their 2D feature map outputs [19]. However, in this work we use global attention mechanism which calculates every time step context vector from the previous time step decoders' hidden state output and the combination of all hidden state outputs of the Lstm encoder (see Figure 4.6). In addition, in this work we implement the proposed approach using both types of Lstm decoders which are one-directional and bi-directional decoders to show the explicit effect of bi-directional Lstm decoder on scene text recognition. However Li et al. [19] uses only one-directional Lstm decoder.

Based on the general architecture of attention based approach as shown in Figure 4.1, we propose an approach which is depicted in Figure 4.2. Detail information about how these four modules are implemented in relation of the proposed approach is given in the following four sub-sections.

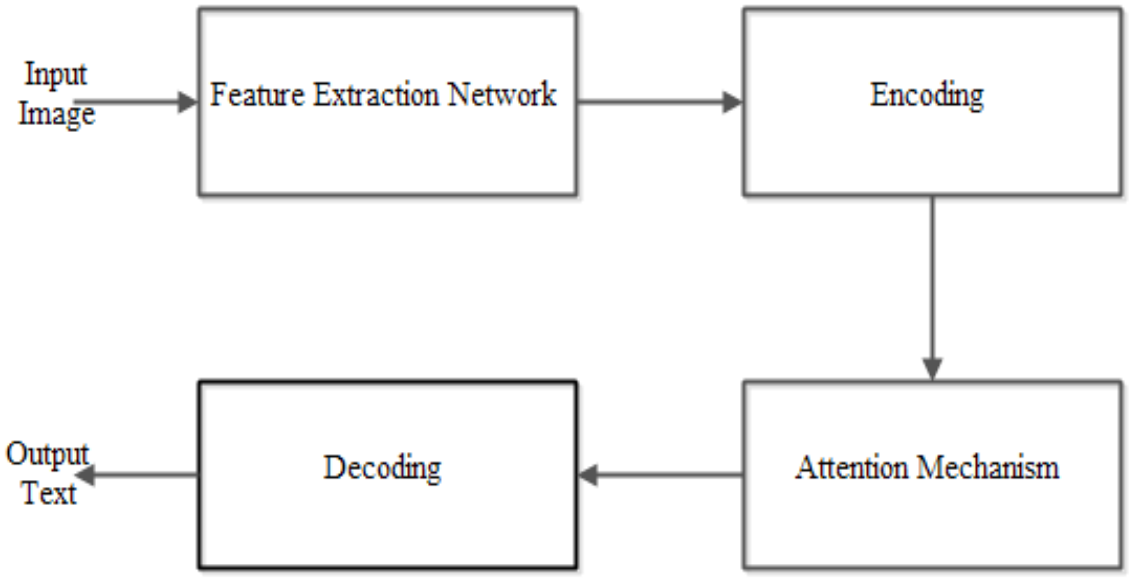


Figure 4.1: General Architecture of Attention-Based Encoder-Decoder Framework in Scene Text Recognition [21]

4.1 Feature Extraction Network

In this module we extract deeper feature representation of scene images by using a ResNet-based CNN. This module maps the input scene image to a representation that focuses on the features relevant for scene text recognition while suppressing irrelevant features such as scene text fonts, scene text colors, scene text sizes, and various scene text backgrounds [28]. As illustrated in Figure 4.2, this module has two main components which are image pre-processing component and ResNet-based CNN component used as a feature extractor.

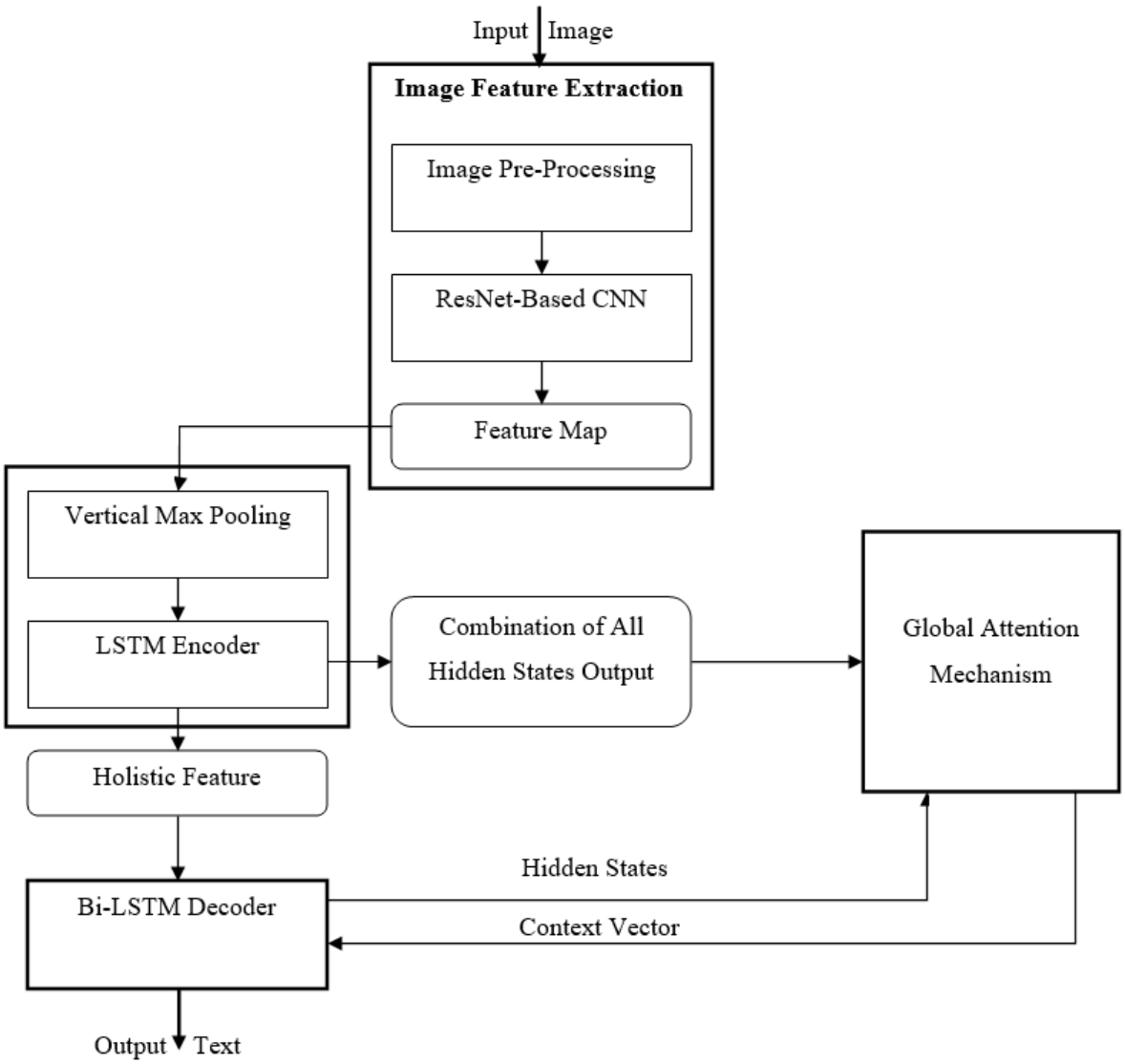


Figure 4.2: Proposed Approach

4.1.1 Image Preprocessing

This component accepts the original scene image as an input and performs preprocessing task which is image rescaling to make the input scene image comfortable for feature extraction stage. Image rescaling improves the accuracy of ResNet based feature extraction and in turn improves the overall recognition performance of attention-based scene text recognition.

Image rescaling performed on MATLAB, since it is very adaptable and comfortable platform for fast and accurate image preprocessing tasks. Preprocessed scene images are used for model

training purpose only. Preprocessed scene images boost up feature learning efficiency of the feature extractor. To test the trained model, we use ICDAR standard evaluation images without any preprocessing.

Training dataset contains scene images of different sizes; image height is in the range of 32 to 46 pixels and image width is in the range of 23 to 256 pixels [34]. When the scene image size is too large, it consumes too much computational cost and processing time. In addition it makes difficult for feature extraction network to learn and extract basic and relevant features [33]. Therefore, the basic reason for rescaling scene images to similar and proportional size is to make the proposed ResNet based feature extractor more robust for relevant feature learning [17].

Cheng et al. [17] rescales all scene images to a fixed size of 32 × 256 to reduce computational time and boosts feature extraction efficiency [17]. Similarly Baek et al. [33] rescales all scene images to a fixed size of 32 × 100. However, when we try to rescale our training datasets with these similar sizes, it becomes difficult to separate individual character sequences available on the image since it is difficult to maintain images aspect ratio. To solve this problem, the baseline work Li et al. [19] resizes scene images to a fixed height of 32 pixels and a varying width to keep their original aspect ratios. Hence, the width of the obtained feature map also varies with respect to aspect ratios. In our work, we also resized scene images to a fixed height of 32 pixels and a varying width to maintain scene texts original aspect ratio similarly as in Li et al. [19].

4.1.2 ResNet-Based CNN Feature Extractor

Feature extraction can be performed using different deep CNN network architectures (see Table 4.2). ResNet is one of those CNN architectures and it is preferable due to its biggest advantage of easy to understand, fast to implement, ease training process and highest accuracy among all CNN network architectures that can predict image features [19], [29], [33].

As a ResNet-based CNN [30] feature extraction module, we implemented the same ResNet-based CNN network which is used in Cheng et al. [17]. The network has 29 trainable layers and 32 layers in total. The detail of the network is shown in Table 4.1.

ResNet-Based CNN feature extractor is widely adopted from the area of object detection and recognition to scene text recognition because of its high representative ability [13]. ResNet is popular and the most powerful network backbone for image feature extraction [2]. ResNet-based network indeed extracts more robust features for scene text recognition and greatly boosts scene text recognition performance [17]. We choose ResNet-based CNN for the proposed feature extraction network since it has high feature extraction accuracy (low error rate), optimized processing time, rational space requirement and deeper representation of scene images [19], [29]. In addition, it avoids the problem of vanishing gradient and boosts up feature learning capabilities [29].

The preprocessed input scene images fed into a ResNet-Based CNN, results in a feature map output, which is a deeper and relevant representation of the scene image. The feature map is ready for encoding stage.

layer name	32 layers	output size
conv1_x	$3 \times 3, 1 \times 1, 1 \times 1, 32$	32×256
	$3 \times 3, 1 \times 1, 1 \times 1, 64$	
conv2_x	pool:$2 \times 2, 2 \times 2, 0 \times 0$	16×128
	$\begin{bmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{bmatrix} \times 1$	
	$3 \times 3, 1 \times 1, 1 \times 1, 128$	
conv3_x	pool:$2 \times 2, 2 \times 2, 0 \times 0$	8×64
	$\begin{bmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{bmatrix} \times 2$	
	$3 \times 3, 1 \times 1, 1 \times 1, 256$	
conv4_x	pool:$2 \times 2, 1 \times 2, 1 \times 0$	4×65
	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 5$	
	$3 \times 3, 1 \times 1, 1 \times 1, 512$	
conv5_x	$\begin{bmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{bmatrix} \times 3$	1×65
	$2 \times 2, 1 \times 2, 1 \times 0, 512$	
	$2 \times 2, 1 \times 1, 0 \times 0, 512$	

Table 4.1: A ResNet-based CNN Architecture for Robust Scene Text Feature Extraction [17]. ResNet building blocks are shown in brackets

Year	CNN	Error Rate (%)	Space Complexity	No of Parameters
1998	LeNet	16.4	1.5 MB	60 thousand
2012	AlexNet	15.3	232 MB	60 million
2014	ZFNet	11.7	255 MB	-
2014	VGGNet	7.3	528 MB	138 million
2015	GoogleNet	6.67	64 MB	4 million
2016	ResNet	3.57	125 MB	25 million

Table 4.2: Comparison of CNN Architectures by Using ImageNet Validation Dataset [29]

4.2 Encoding

After feature extraction stage, the extracted features of the input scene image are encoded into a new binary feature sequence representation with an encoding module. Encoding module converts the feature map into a new, intermediate and abstract sequential representation which can be understood by both the encoder and the decoder [19]. Therefore, encoders and decoders are used to convert data from one form to another form [20]. Encoding module is a crucial module in attention-based encoder-decoder scene text recognition system. The decoder cannot understand the feature map representation without encoding module. Therefore, the feature map has to be encoded to a new and abstract one-dimensional binary feature representation that can be understood by the decoder for correct character sequence generation. As illustrated in Figure 4.2 this module has two main components which are vertical max pooling and LSTM Encoder.

4.2.1 Vertical Max Pooling

Pooling layer is used as a refinement module to reduce the dimension of the feature map. By summarizing the features present in a specific region of the feature map, it reduces the number of parameters to learn and the number of computations performed in the model [33]. Vertical Max Pooling is a type of pooling operation that calculates the maximum or largest values in each vertical patch of the given feature map. As we see from Figure 4.3, it produces down-sampled feature maps that highlights the most present features in the vertical patch of the original feature maps. By reducing the total number of parameters in the model, it also minimizes the chance of occurrence

of over fitting [19]. So that, at each time step of encoding, a column feature is firstly max-pooled along the vertical direction, then after it passes to LSTM encoder. In this work for vertical max-pooling operation, we use stride of 2 and a filter size of 1 × 2 in a vertical direction which totally reduces the dimension of the original feature map by a factor of 2.

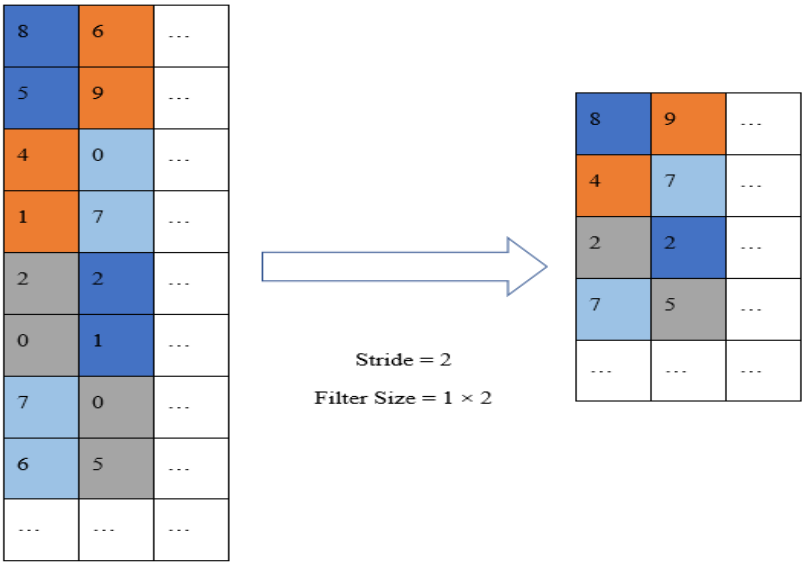


Figure 4.3: Clarification of How our Vertical Max Pooling Operation Works. Original Features at the Left Side are Down-Sampled to Features at the Right Side with Similar Colors

4.2.2 LSTM Encoder

Long and Short-Term Memory networks usually just called ‘LSTMs’ are a special kind of Recurrent Neural Network (RNN), capable of learning long-term and short-term pattern dependencies [53]. They have an excellent ability of learning sequential data (feature patterns) [1]. Remembering information for long periods of time is practically their default behavior. Due to this unique behavior, LSTMs are enabling recognition of more than one-character strings on the scene image sequentially as a word text string. In this work, LSTM encoder encodes the down-sampled feature map sequentially, column by column, and finally produces the holistic feature as an output. This holistic feature used as an initial input for the decoder. The general architecture of the LSTM encoder shown in Figure 4.4 is for the encoding module.

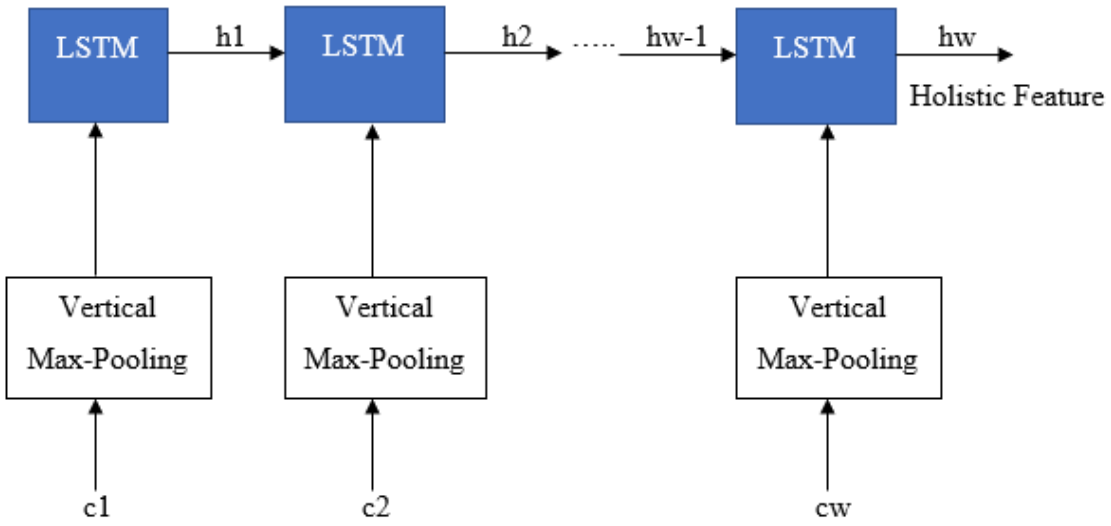


Figure 4.4: The Structure of the Encoding Module. cj represents the jth Column of the Feature Map

In this work as an encoding module, we adopt a 2-layer LSTM model with 512 hidden state size per layer which is proposed by Li et al. [19]. Duplicating number of LSTM layers is used to enhance model capacity and improve overall model accuracy for sequence-to-sequence learning problems [31]. When the number of layers increase, the model capacity increases along with accuracy. However, since its computational cost and time also becomes higher and higher, the number of layers must be optimized based on available resources. At each time step t, the LSTM encoder receives one column of the feature map after max-pooled along the vertical axis, and updates its hidden state to ht which is the current hidden state at time step t. Encoding process is performed iteratively from column one (c1) up to the width size (cw) of the feature map. After w steps, the final hidden state of the second LSTM layer, which is 'hw', is regarded as a fixed-size holistic feature representation of the input image, and provided for decoding module.

4.3 Attention Mechanism

Attention mechanism is a powerful tool, which is developed to enhance the performance (task-specific accuracy) of the encoder-decoder model in neural network-based tasks [26]. Therefore, it adds a flavor for neural network-based tasks which needs an explicit focusing mechanism. Attention mechanisms are applied to a neural network by which a network can weigh different

patterns and features by their level of importance to a task, and use this weighting to help to achieve the task with greater performance [25]. The task can be recognition, classification, detection, translation or captioning. In attention-based encoder decoder framework, attention is used as an interface that connects the encoder and the decoder to facilitate exchange of valuable information. It helps the framework to achieve a given task with a higher focus and accuracy [26], [33].

The concept of attention in neural network-based encoder-decoder framework is analogous with visual attention in humans [46]. Visual attention is an important part of everyday life [26]. Focusing on the road while we are driving, glancing at the food on our plate before we take a bite, and looking at the text instead of the sidebars when we are reading are example scenarios of visual attention. However, when we are training a neural network, we need a way for it to explicitly incorporate the concept of visual attention. In order to make correct decisions the network needs to be able to learn the most important and relevant parts to focus on [25]. In this work as an attention module, we proposed global attention mechanism which attends all the input characters during each decoding time steps instead of attending a subset of characters in the case of local attention mechanisms.

4.3.1 Global Attention Mechanism

Sequence-to-sequence models have been widely used in machine translation, speech recognition, image captioning, video captioning, time serious forecasting, text summarization and text recognition [31], [32], [17]. The problem we are addressing also falls under sequence-to-sequence problem, hence we adopted the concept of global attention in the area of speech recognition [32] and image captioning [46] to our encoder-decoder framework of scene text recognition. Without explicitly transforming distorted scene images, the proposed global attention module is able to accommodate text of arbitrary shapes, layouts, distortions and orientations implicitly.

As described in Section 3.4, long sequences are outperformed by the current state-of-the-art approach, Li et al. [19] due to poor encoder output utilization. It only utilizes the last hidden state output of the encoder holistically only once. When we go from the first time step of the decoder to the last time step of the decoder, the connection between the holistic feature and decoder outputs

becomes weaker and weaker. The connection between encoders' holistic feature and the third decoder output is better than the fourth decoder output, and also the connection to the fourth decoder output is better than the fifth decoder output, and so on for all decoder output sequences until the end. Therefore, it is very far away for those decoder outputs at the end of the sequence to extract and use the holistic feature information to produce the correct output text.

In this work, unlike [19] we use not only the holistic feature of the encoder but also, the combination of all the encoder hidden state outputs generated at all consecutive time steps. Unlike Li et al. [19], the proposed global attention mechanism calculates the context vector (g) from the decoders' previous time step output and the combination of all encoder hidden state outputs as seen in Figure 4.5, instead of calculating from the feature map for all decoding time steps in the baseline work. By doing this we can handle attention mismatch problem which is misalignment between feature areas and target areas for irregular and distorted texts.

Figure 4.5 (which is adapted from https://www.colab.research.google.com) shows the detail architectural configuration of global attention mechanism with encoder and decoder framework.

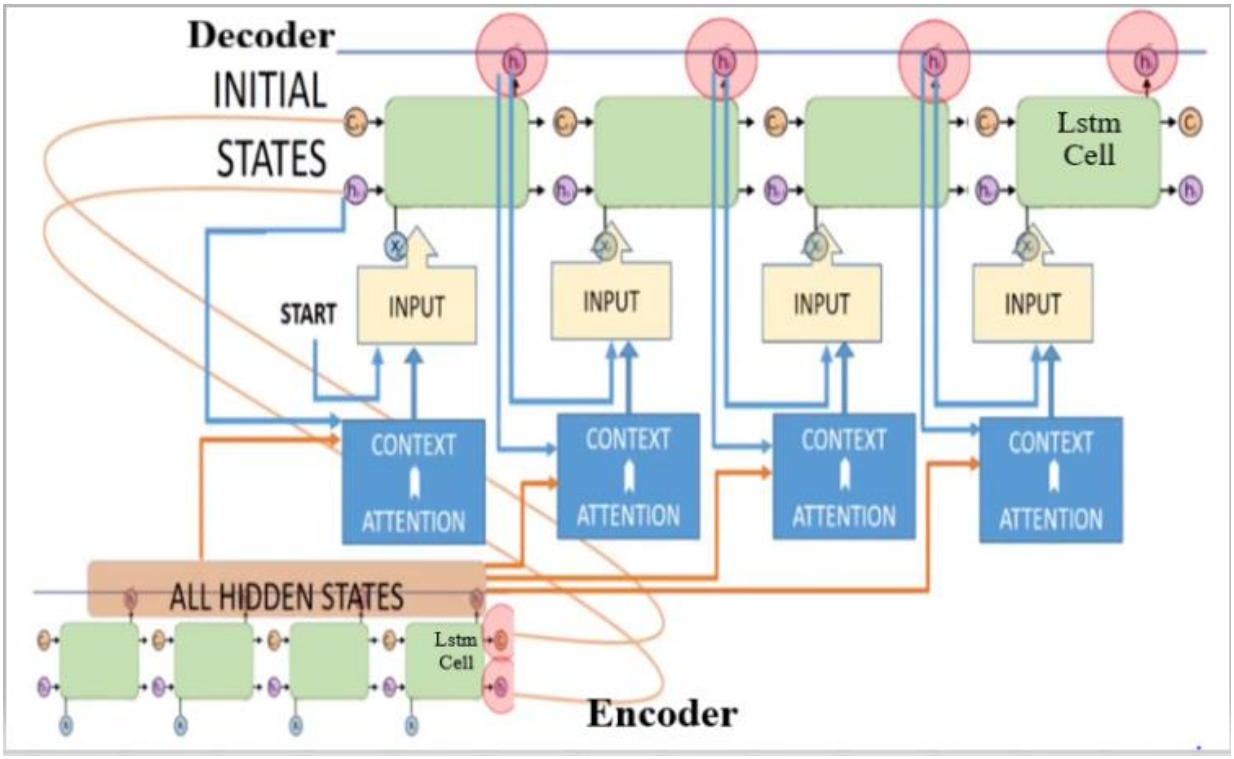


Figure 4.5: Architectural Configuration of Global Attention Mechanism with Encoder and Decoder

At the beginning, we initialize the decoder states by using encoders' holistic feature (hw) as an initial input then at each decoding time step, we use combination of all encoder hidden state outputs and the decoder's previous time step output to calculate the context vector by applying a global attention mechanism. This means, we have to attend all the encoder outputs for generating each time step decoders' output as illustrated in Figure 4.6. However, in the case of local attention mechanism we have to attend (consider) only a subset of encoder outputs for generating each time step decoder outputs.

Figure 4.6 (which is adapted from https://www.theaisummer.com) shows the basic difference between global and local attention mechanisms in an abstract view for better understanding of the working principles behind them.

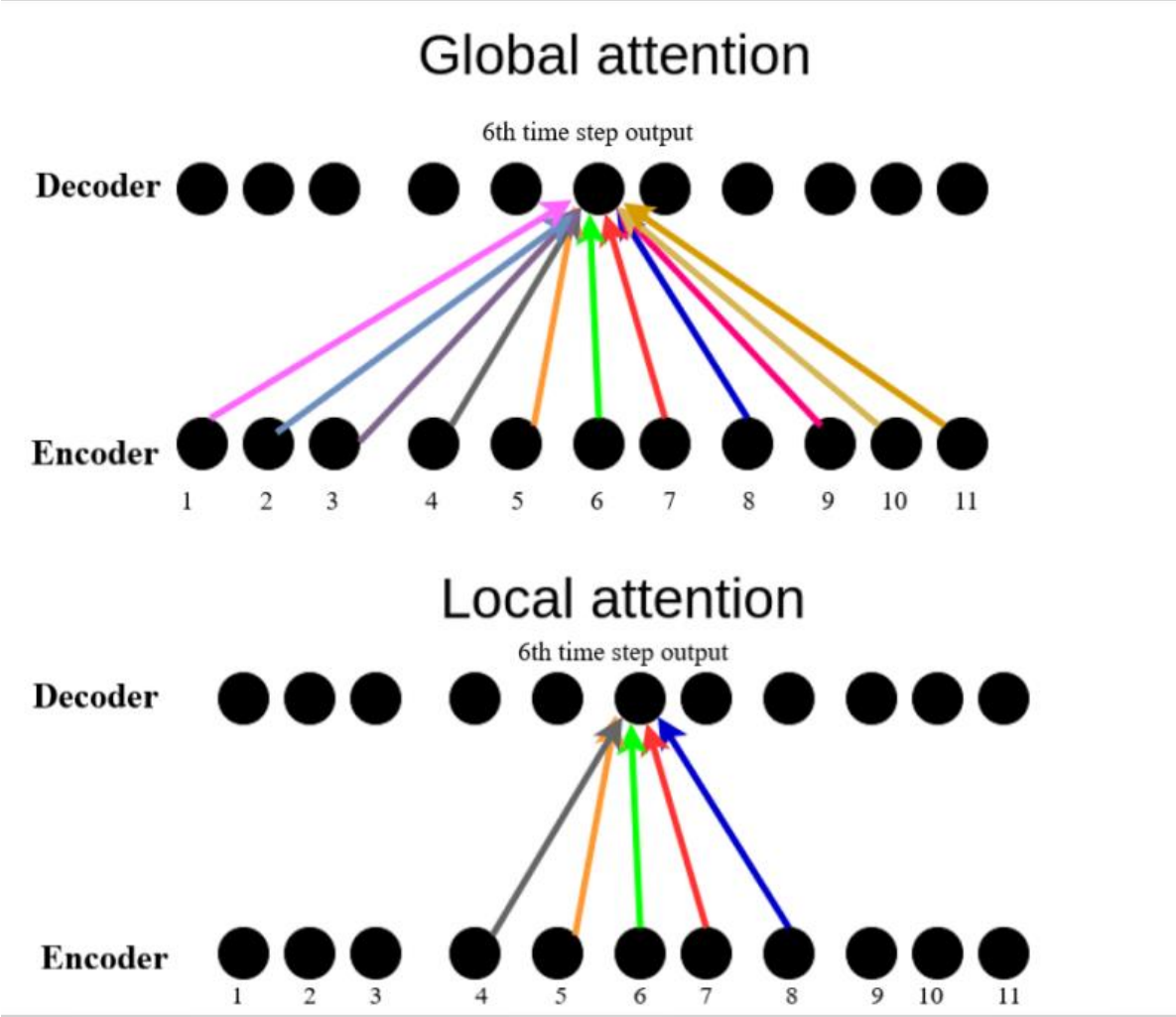


Figure 4.6: Demonstration of how Global and Local Attention Mechanism Works

Local attention can also be merely seen as hard attention since we need to take a hard decision first, to exclude some encoder output units [56]. The colors in both attention mechanism indicate that these weights are constantly changing.

Global attention usually eliminates the vanishing gradient problem, as it provide direct connections between the encoder states and the decoder [27], [56]. Conceptually, it act similarly as skip connections in convolutional neural networks [56].

Calculating the context vector in global attention mechanism involves three step processes. First calculate the score, then using the score calculate attention weights, finally calculate the context vector from attention weights.

Below we describe the steps followed to calculate the context vector in detail.

- Notations

h_s : Combination of all hidden state outputs of the encoder

h_t : Previous time step output of the decoder

g_t : Context vector of time step t

W : Weight matrix for parametrizing the calculations

- Step 1: Calculate the score to relate combination of all hidden state outputs of the encoder and the previous time step output of the decoder by using bahdanau's additive score formula [56] using Equation 4.1.

$$Score(h_t, h_s) = V^T \tanh(W_1 h_t + W_2 h_s) \quad (4.1)$$

The score determines the level of relationship between the combination of encoders' all hidden state outputs and the previous time step decoders' output. We use 'W' matrices to parametrize the calculations. The correct weight values are determined practically during training via backpropagation. Therefore, the model will learn how to calculate better scores. Trainable parameters W, V and tanh are a single hidden layer network models [33]. As a result, we expect that these layers W, V, and tanh will learn how to calculate a suitable score during training.

- Step 2: Calculate the attention weights or alignment factors as in [17] by normalizing the scores as we see from Equation 4.2.

$$\alpha_{ts} = \frac{exp\big(Score(h_t, h_s)\big)}{\sum_{s'=1}^{s} exp\big(Score(h_t, h_{s'})\big)} \tag{4.2}$$

Where s′ represents a time step. It starts from 1 for the first time step and increment iteratively until it becomes 's' (the window size) which is the last time step value [56].

Attention weights (α) are the weights for each encoder hidden state output in h_s. When we sum up the attention weights of all encoder hidden states, we get 1.0 [32]. The weights indicate which hidden state output in the encoder is important and by how much with respect to the last output in the decoder [17].

- Step 3: Calculate the context vector by applying the attention weights onto all encoder hidden state outputs ($h_{s)}$ using Equation 4.3. Thus, we will have weighted encoder hidden state outputs (g_t) at the end.

Using attention weight distributions, we can multiply encoder hidden state outputs with attention weights. This helps to strengthen the important one and weaken the least important one as a context vector. The decoder then can decides parts of the input feature sequences to pay attention to.

$$g_t = \sum_{s} \alpha_{ts} h_s \tag{4.3}$$

Generally the idea behind attention mechanism is the decoder is often trained to predict the next character given the context vector and all the previously predicted characters. The context vector calculation (representation) varies for different types of attention mechanisms [32]. In this work we use global attention mechanism which considers all hidden state output of the encoder to generate every time step decoders' output. Hence, to calculate the context vector first it try to relate the previous hidden state output of the decoder with every time step hidden state output of the encoder by applying SoftMax normalization to produce a scalar value as an attention vector for all encoder hidden state outputs. Then multiply this attention vector value with values of each encoder

hidden state output to get alignment vector value. Finally by summing up alignment vector values of every time step in s we get a context vector [56].

4.4 Decoding

After encoding operation, the generated holistic features are decoded into a sequence of characters using a decoding process of an LSTM decoder. The decoder retrieves the enriched abstract representation of sequential features from the encoder to generate a sequence of characters [27]. The general structure of the LSTM decoder used in this work is shown in Figure 4.7.

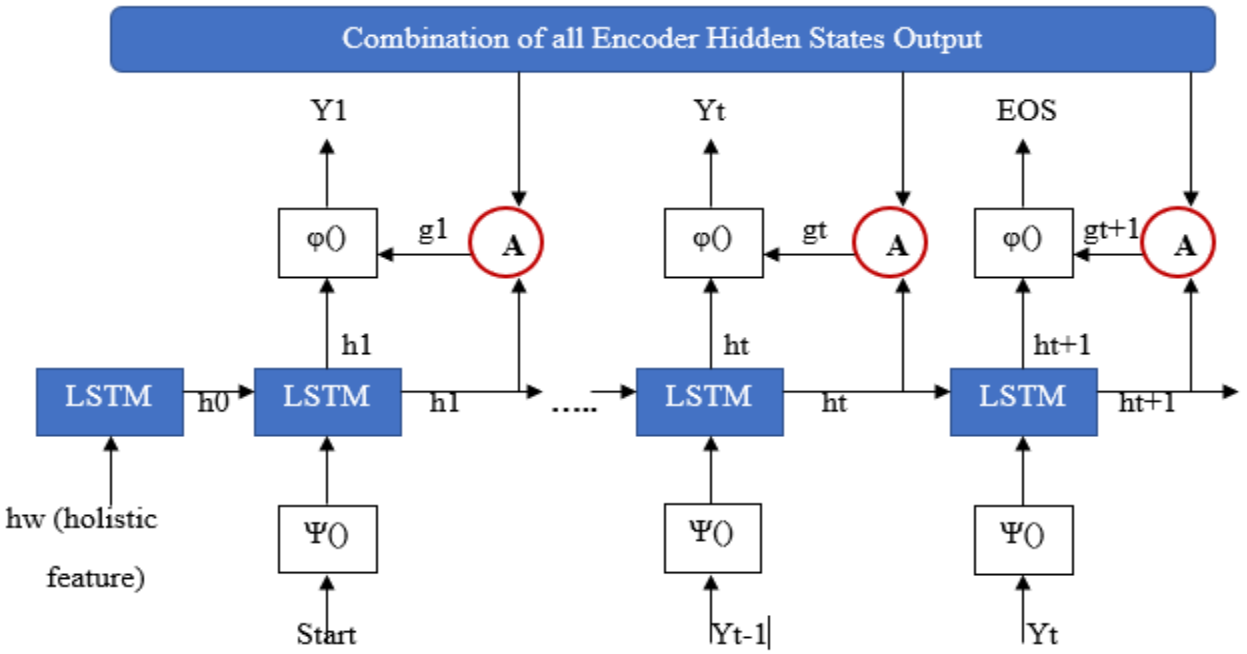


Figure 4.7: The Structure of the Decoding Module. **A**: represents Global Attention Module

The decoder is a 2-layer LSTM model with 512 hidden state size per layer. The holistic feature (hw), a Start token and the previous outputs are input into LSTM subsequently, terminated by an EOS (end of string) token. In order to train an LSTM neural network to generate text, we must first preprocess our text data so that it can be consumed by the network. In this case, since a neural network like LSTM takes only vectors as an input, we need a way to convert the text into vectors. All inputs to LSTM are represented by one-hot vectors, which are the converted one-hot-encoded-format of 37 alpha-numeric character classes which are 26 case-insensitive alphabet characters, 10

digits and 1 EOS token by a linear transformation function Ψ() which performs one-hot-encoding operation. In this case a function Ψ() produces a one-hot-vector representation which is a 1 × 37 matrix (vector) used to distinguish each 36 alpha-numeric character and EOS token from every other alpha-numeric character. The vector consists of zeros in all cells with the exception of a single 1 in a cell used uniquely to identify that unique alpha-numeric character. During training, the inputs of decoder LSTMs as a previous time step output are replaced by the ground-truth character sequences, which comes from word level annotations.

At each time step t, the output (Y_t) is computed by the function φ() with the current hidden state (h_t) and the current context vector (g_t) as an input. The current context vector (g_t) is calculated by the global attention module as shown in Equation 4.3. Finally, the outputs are computed by applying the transformation function in Equation 4.4.

$$Y_t = \varphi(h_t, g_t) = SoftMax(W_o[h_t; g_t]) \quad (4.4)$$

Where W_o is a linear transformation, which embeds features into the output space of 37 classes, in corresponding to 26 English alphabet letters (case insensitive), 10 digits and an EOS token.

The LSTM decoder works only in one direction which is from left to right. This, however, may recognize some characters incorrectly. For example, in some font types; letter B may be incorrectly recognized as letter E or letter R, letter F may be incorrectly recognized as letter P, letter J may be incorrectly recognized as letter I, letter Q may be incorrectly recognized as letter O, letter S may be incorrectly recognized as number 5, etc. Therefore, in order to minimize this problem and to leverage the dependencies in both directions, we propose a bidirectional, Bi-LSTM decoder which works in both directions as in Shi et al. [1].

4.4.1 Bi-LSTM Decoder

Although a sequence-to-sequence decoder captures output dependencies, it captures the dependency only in one direction and misses the other. For example, a decoder that recognizes text from left-to-right order may have difficulty to decide the character either upper-case ‘I’ or number ‘1’ in certain fonts, as they are hard to distinguish visually. Such texts may be easily recognized by a decoder that works in the right-to-left direction [1]. As this example suggests,

decoders that work in opposite directions are potentially complementary. One decoder is trained to predict characters from left to right and the other from right to left. Two independent recognition results are produced after running both LSTM decoders as illustrated in Figure 4.8. To merge the results, we simply pick the output of the decoder with the highest recognition accuracy as bidirectional decoder output. However, using PyTorch [54] package, we can implement both decoders as one Bi-directional LSTM decoder. The total loss of the Bi-LSTM decoder is the average of the losses of the two LSTM decoders and calculated as in Equation 4.5.

$$L_{total} = -\frac{1}{2}\sum_{t}\left(\ln P_{l2r}\left(Y_t|I\right) + \ln P_{r2l}\left(Y_t|I\right)\right) \tag{4.5}$$

Where I represent the actual input scene text on the image, Y_t is the predicted output text and P is the prediction operation.

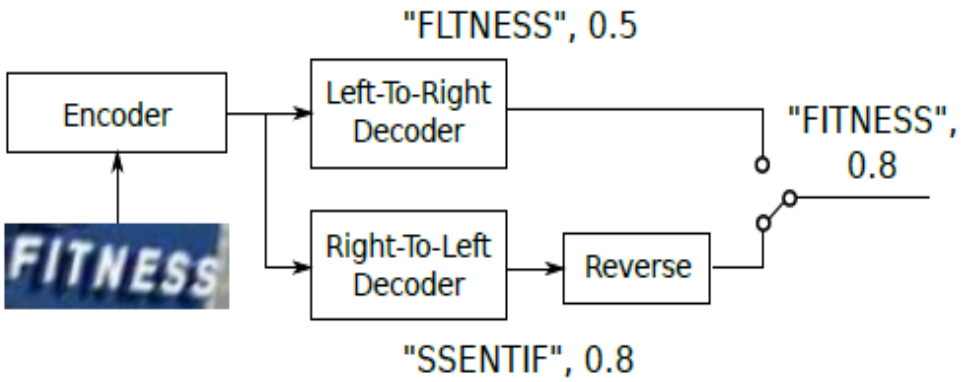


Figure 4.8: Bi-LSTM Decoder [1]. '0.5' and '0.8'are example recognition accuracies

4.5 Performance Evaluation Metrics

In this section, we present the evaluation metrics need to assess scene text recognition (STR) performance. The metrics need for assessment is average scene text recognition accuracy. We measure the success rate of scene text recognition accuracies for all scene images on the 7 class of real-world scene text recognition model evaluation datasets involving 4 class of regular scene text datasets and 3 class of irregular scene text datasets. A unified 7 class of evaluation dataset consists of 8,374 scene images in total: 3,000 from IIIT5K, 647 from SVT, 860 from ICDAR 2003, 857 from ICDAR 2013, 2,077 from ICDAR 2015, 645 from SVT-P, and 288 from CUTE80. We only

evaluate the proposed approach on all English alphabets and digits which are 26 case-insensitive alphabet characters and 10 digits which become a total of 36 alpha-numeric character combinations. Hence, we have not considered recognition of special symbols and punctuation marks.

Average scene text recognition accuracy of each scene text image is measured by the weighted sum of each of the decoded alphanumeric character's confidence score [55]. A Confidence Score is a number between 0 and 1 (100%), that represents the probability of the decoded output of a model is belonging to 37 output classes (26 English alphabet letters, 10 digits and 1 End of string token). A confidence score of 1 is likely an exact match for that specific class. The total sum of all confidence score is 1 for every decoding time step.

For each decoding time step, our model predicts the likelihood of the decoded output for belonging to 37 class of output spaces and assigning conditional probabilities for all of them by using the SoftMax activation function as described in Equation 4.4. The higher the conditional probability means the greater the confidence score that the decoded output to fall in that specific class. To calculate the confidence score, we utilize the conditional probabilities assigned to the first two most probable class of output spaces by using greedy search.

At decoding time step of t, let Y_{t1} and Y_{t2} be, the two most likely embedding class of output space, and Pr $[Y_{t1}|Y_t]$ and Pr $[Y_{t2}|Y_t]$ are their conditional probabilities. Then the confidence score for Y_{t1} which has the highest conditional probability among all predictions can be calculated by using Equation 4.6 and it is displayed as the final decoded alpha-numeric character.

$$Confidence\ Score\ Per\ Character = 1 - \frac{\Pr[Y_{t2}|Y_t]}{\Pr[Y_{t1}|Y_t]} \tag{4.6}$$

These character confidence scores for all decoding time steps are then weighed and summed to obtain confidence score of text strings as in Equation 4.7, which serves as the overall confidence score for the text string on the scene image.

$$Confidence\ Score\ Per\ Image\ = \frac{\sum_1^N (Confidence\ Score\ Per\ Character)}{N} \tag{4.7}$$

Where N is total number of decoded characters.

The image confidence scores are then weighed and summed to obtain confidence score of each dataset classes as in Equation 4.8, which serves as the average recognition accuracy for a particular dataset class.

$$Average\ Recognition\ Accuracy\ = \frac{\sum_1^N (Confidence\ Score\ Per\ Image)}{N} \tag{4.8}$$

Where N is total number of scene text images per each dataset classes.

Chapter 5

Experiment and Results

In this Chapter datasets, tools and experimental setups used to evaluate the proposed approach along with the results obtained are discussed. The proposed approach is evaluated based on the performance evaluation metrics discussed in Section 4.5.

5.1 Experimental Setup

All modules in the proposed two-dimensional global attention-based scene text recognition system are implemented in Python 3.8.5 using version 1.8.1 of Torch, version 0.9.1 of TorchVision, version 1.19.2 of NumPy, version 2.5.0 of TensorFlow and version 2.4.3 of Keras. These are Python packages for deep learning-based tasks with image and video processing. They provide implementations of a variety of common functions used throughout the field of text detection and recognition, particularly for scene text detection and recognition from natural scene images. The hardware and software specification of the machine, that is used to implement the proposed system and conduct the experiments is given in Table 5.2.

5.2 Dataset Description

In this section we describe about the training and evaluation datasets used in the experiments. The proposed approach is trained using SYN90K synthetic dataset and evaluated under 7 class of scene image datasets which are 4 class of scene images with regular scene text and 3 class of scene images with irregular scene text as shown in Table 5.1. The training dataset contains 6,000 randomly selected regular and irregular scene text images from SYN90K. For evaluation, we need scene text images under 7 classes with 2 categories (see Table 5.1). The evaluation scene images are commonly used for testing any type of scene text recognition models [1].

5.2.1 Training Datasets

There are over hundreds of image dataset classes available for image-based researches. These datasets are distinct in number, feature and characteristics [43]. Since, almost all of these image dataset classes are not uniform and not applicable throughout the world with similar standards, it is difficult to compare different models relied on them [33].

For researchers working on image-based computer vision tasks, there are 3 publicly available, large scale, general and uniform class of image datasets for training purpose and as well as for evaluation purpose in some works. These are SYN90K [34], SynthText [35] and ImageNet [43] dataset classes. Below is the description of the three general classes of datasets.

In this work, we use SYN90K synthetic datasets [1] for training purpose, mainly because it is designed for end-to-end training of scene text recognition tasks. This dataset class is synthetic in nature. Therefore, it was not used for validation and evaluation purposes.

For this work, all of the training scene image datasets are preprocessed according to the preprocessing tasks discussed in Section 4.1.1. All scene images are resized and filtered for training the model as shown in Figure 5.1 as a sample.

1) **SYN90K:** is the synthetic scene text dataset [34]. This dataset class contains over 9 million synthetic scene images generated from a collection of 90 thousand common English words [19]. Words are rendered onto natural images with random transformations and effects. Every image in this class is annotated with a ground truth word level annotation. All of the images in this dataset class have been widely used for training of scene text recognition models.
2) **SynthText:** is a synthetic and large-scale scene text image datasets proposed by [35]. This dataset class contains over 8 million synthetic word images. It was particularly created for scene text detection researches [1], [2]. It has been widely used for scene text recognition research by cropping text image patches using the ground truth word bounding boxes [23]. Researchers can crop and use different amounts for evaluations depending on their need.

[1] https://www.robots.ox.ac.uk/~vgg/data/text

3) **ImageNet:** is a large-scale image database [43]. It is built upon the hierarchical structure provided by WordNet [44]. Each meaningful concept in WordNet, described by multiple words or word phrases, is called a Synonym Set or (SynSet). This dataset class contains over 14 million images with 22,000 visual categories [43]. It has been widely used for object detection, object classification and object recognition researches.

Figure 5.1: Sample Scene Text Images from the Training Dataset after Preprocessing

5.2.2 Evaluation Datasets

The proposed approach is evaluated on 7 scene text image dataset classes, similar to other related works [1], [2], [17], [19], [20] on scene text recognition. The dataset includes 4 regular scene text dataset classes which are ICDAR 2003 (IC03), ICDAR 2013 (IC13), IIIT5K, and SVT. Most of scene texts in these classes are almost horizontal and frontal parallel [33]. The other 3 irregular scene text dataset classes are ICDAR 2015 (IC15), SVT-Perspective (SVT-P) and CUTE80 (CT80). In these 3 datasets, a large amount of scene texts suffers from perspective and curvature distortions. The 7 class of scene text image datasets are publicly available and have been widely used for evaluations in almost all scene text recognition researches [1], [17], [19]. Detail

information about these 7 evaluation dataset classes is given below and their samples are illustrated in Figure 5.2. All of these 7 classes of scene text datasets are available here in [2].

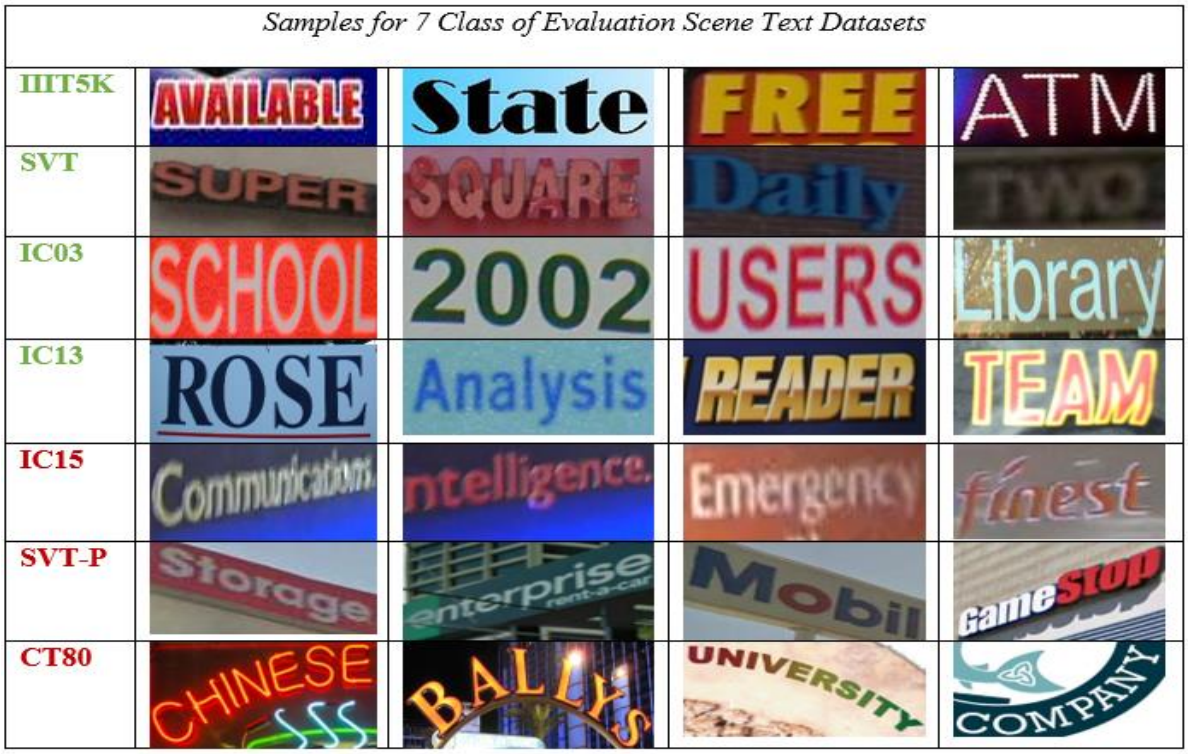


Figure 5.2: Sample Scene Text Images from Evaluation Dataset Classes

1) **ICDAR 2003 (IC03)** [36] is used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2003. It contains 860 scene text images of cropped words after filtering [40]. The filtering discards words that contain non-alphanumeric characters or have less than three characters. Scene texts in this class are horizontal and frontal parallel.
2) **ICDAR 2013 (IC13)** [37] is used in the Robust Reading Competition in the International Conference on Document Analysis and Recognition (ICDAR) 2013. It contains 857 scene text word images for testing purpose. All scene text images contain text in a variety of colors and fonts on various backgrounds [1]. Most scene texts in this class are horizontal.

[2] https://github.com/bgshih/aster/releases/tag/v1.0.2

3) **ICDAR 2015 (IC15)** [38] is a very popular benchmark used in the Robust Reading Competition under ICDAR 2015. It contains incidental scene text images that are captured without preparation and camera adjustment before capturing [19]. Generally, 2077 scene text image patches are cropped from this dataset class, where a large amount of cropped scene texts suffer from perspective and curvature distortions. It also contains hundredth of scene text images with extremely low resolution.
4) **IIIT5K** [39] consists of 3000 test scene text images that are cropped from natural scene images which are collected from the web by google image search. Scene text instances in these images are nearly horizontal but distinct in font, color, style, and background [39].
5) **Street View Text (SVT)** [40] is collected from the google street view images that were used for scene text detection research. Generally, 647 scene text word images are cropped from 249 street view images and most words within this cropped scene text image classes are almost horizontal.
6) **SVT-Perspective (SVT-P)** [41] consists of 645 scene text images that are cropped from the street view text images. Most scene text images in this dataset class are heavily distorted by perspective distortions which are specifically picked for evaluation purpose of scene text recognition models under perspective views [1], [19].
7) **Curved Text-80 (CUTE80/CT80)** [42] consists of 288 scene text images where most cropped words are curved in different angles. All word level scene text images are cropped from the CUTE dataset which contains 80 high resolution scene text images that are originally collected for evaluating the performance of curved text recognition of different scene text recognizers [33].

Most approaches particularly approach without attention mechanisms use lexicons [20]. A lexicon consists of the ground truth word and other random words as a collective and finite word library. If lexicons are available, during testing a model the probability distributions for all words are evaluated and the word with the highest probability chosen as the final constrained result [3]. Lexicons are mostly important for scene text recognition models with bounding boxes [15]. However, in this work, the proposed approach works in a lexicon-free mode. Therefore, the decoder outputs the predicted sequence of texts in unconstrained manner.

	Scene text image dataset classes	Total number of scene text images
Evaluation datasets for Regular scene text	IIIT5K	3,000
	SVT	647
	IC03	860
	IC13	857
Evaluation datasets for Irregular scene text	IC15	2,077
	SVT-P	645
	CT80	288

Table 5.1: Evaluation Scene Text Datasets with their Total Number of Scene Text Images

5.3 Experimental Scenarios

Based on their intended goal, two experiments are conducted to evaluate the proposed scene text recognition system. We conduct both experiments under the same training parameters except one directional LSTM decoder in experiment I is substituted with two directional Bi-LSTM decoders in experiment II. The proposed system has been trained for 50 epochs and it approximately takes 7 days to finish the training process. We used Adam as an optimizer since it has fast convergence speed. The learning rate is 0.001 which is the default value for Adam optimizer. We used batch size of 8. Therefore, there are 750 training iterations under each epoch.

Epoch is the number of passes of the whole training dataset through the network. The training dataset is divided into 8 batches and iterated through the network, one batch at a given period of time. When all the batches are iterated and the whole training dataset has gone through the network completely, this is called one epoch [21].

- **Experiment I:** Aims to evaluate the proposed global attention mechanism on scene text recognition performance.

- **Experiment II:** Aims to see the explicit effect of using bi-directional LSTM decoder (Bi-LSTM decoder**)** over one directional LSTM decoder (LSTM decoder) on scene text recognition performance. Some scholars say that Bi-LSTM decoder increases model complexity without any significant performance improvement. However, others argued that it improves scene text recognition performance. Therefore, we need a way to show Bi-LSTM's explicit effect.

In experiment I, the proposed global attention-based scene text recognition system is compared with the previous state-of-the-art local attention-based encoder decoder framework SAR as the baseline work. For the baseline work, SAR, our own reimplementation based on an open-source work presented by Li et al. [19] is used. The source code for SAR is available on GitHub, https://github.com/liuch37/sar-pytorch, under MIT License.

Machine Specification Used for Experiments	
Manufacturer	HP
Model	HP OMEN
Processor	Intel R Core (TM) i7- 10th Generation
Graphics (GPU)	GeForce RTX 2080 SUPER
Memory	8 GB GDDR6
SSD + HDD	256 GB + 1000 GB
Operating System	Windows 10 Home 64

Table 5.2: Machine Specification

5.4 Experimental Results

In this section we provide the results along with their brief explanations. The results are depicted based on the order of the experiments and the discussion part in Section 5.5 summarizes all the results and explains them in relation with the research questions mentioned in Section 1.1.

5.4.1 Experiment I: Evaluation of Proposed Approach

Experiments under this category aim to evaluate the ability of the proposed approach to recognize scene texts from the given scene images accurately regardless of scene text orientations whether regular or distorted (irregular). This experiment comprises 7 sub experiments for all 7 class of evaluation datasets which are 4 class of regular test datasets and 3 class of irregular test datasets as described in Table 5.1.

As we see from Table 5.3, the proposed approach outperforms the state-of-the-art approach. When we compared with the baseline approach Li et al. [19], the proposed approach shows significant improvements both in regular and irregular scene text recognition tasks in all class of evaluation benchmarks except for IIIT5K and IC15 benchmarks, whose recognition results are indicated on gray shades in Table 5.3.

		Recognition Accuracy (%)	
		Proposed Approach	Li et al. [19]
Regular Test Dataset	IIIT5K	64.28	65.05
	SVT	63.53	59.80
	IC03	69.36	66.41
	IC13	65.55	65.13
Irregular Test Dataset	IC15	52.25	52.64
	SVT-P	56.96	53.61
	CT80	53.17	50.58

Table 5.3: Results for Experiment I: Comparisons of Scene Text Recognition Accuracy with Baseline Approach on 7 Benchmarks

5.4.2 Experiment II: Effect of Bi-LSTM Decoder

This experiment aimed to see the explicit effect of Bi-LSTM decoder in the proposed approach. In experiment I, we evaluate scene text recognition performance of the proposed approach with one directional LSTM decoder under all benchmarks and compare it with the state-of-the-art baseline approach. However, in this experiment, we substitute LSTM decoder of the proposed approach in experiment I with two directional Bi-LSTM decoder and compare the results to know the explicit effect of Bi-LSTM decoder based on its contribution on scene text recognition performance gains.

As we see from Table 5.4, the proposed approach with Bi-LSTM decoder outperforms the proposed approach with one directional LSTM decoder. By recognizing scene texts in both directions from left to right and from right to left, the Bi-LSTM decoder contributes a significant effect on the overall scene text recognition performance of the proposed approach. Table 5.3 and Table 5.4 show that the proposed approach outperforms the state-of-the-art in all cases while using Bi-LSTM decoder. Therefore, we can conclude that the proposed approach is robust for all kind of distortions, as it outperforms the state-of-the-art approach in almost all cases while using both types of LSTM decoders.

		Recognition Accuracy (%)	
		Proposed Approach	Proposed Approach with Bi-LSTM Decoder
Regular Test Dataset	IIIT5K	64.28	71.01
	SVT	63.53	70.08
	IC03	69.36	71.27
	IC13	65.55	71.34
Irregular Test Dataset	IC15	52.25	57.38
	SVT-P	56.96	60.70
	CT80	53.17	53.45

Table 5.4: Results for Experiment II: Comparisons of Scene Text Recognition Accuracy for Proposed Approach with Bi-LSTM Decoder and with One Directional LSTM Decoder

5.5 Discussion

In this section we summarize all collected results from the experiments and explain them in relation to the research questions presented in section 1.1.

RQ1 [Global Attention Mechanism]

Can we improve the overall scene text recognition performance by using global attention mechanism in regular and irregular scene text recognition?

The results obtained from experiment I can answer RQ1. From this experiment we demonstrated that by using global attention mechanism, the proposed approach improves the overall performance of regular scene text recognition by an average of 1.58% and irregular scene text recognition by an average of 1.85% (see Table 5.5). In addition of quantitative experimental results, we can see qualitatively as illustrated in Figure 5.4, the proposed approach performs better than the state-of-the-art baseline approach in both regular and irregular type of perspective and curvature distortions of scene texts. The improvement is mainly due to the incorporation of global attention mechanism.

	Recognition accuracy (%)	
	Regular Test Dataset	Irregular Test Dataset
Proposed Approach	65.68	54.13
Proposed Approach with Bi-LSTM decoder	70.93	57.18
Baseline Approach	64.10	52.28

Table 5.5: Comparison of Overall Average Scene Text Recognition Accuracy for both Regular and Irregular Test Datasets

RQ2 [Bi-LSTM Decoder]

What is the effect of using Bi-LSTM decoder instead of using one directional LSTM decoder on the performance of regular and irregular scene text recognition?

The results obtained from experiment II can answer RQ2. From this experiment, we demonstrated that by using Bi-LSTM decoder instead of using one directional LSTM decoder, the proposed approach gains an average of 5.24% overall performance improvement on regular scene text recognition and an average of 3.05% overall performance improvement on irregular scene text recognition. Therefore, using Bi-LSTM decoder instead of one directional LSTM decoder gives a positive enhancement effect in the proposed approach both in recognition of regular and irregular scene texts.

As we can see from Figure 5.3 of all model's performance comparisons, the large gain of scene text recognition performance is obtained from the incorporation of Bi-LSTM decoder in the proposed approach. However, without the use of Bi-LSTM decoder the proposed approach also outperforms the state-of-the-art approach both in recognition of regular and irregular scene texts in 5 of 7 cases. It outperforms the baseline approach on recognition of regular scene text datasets by an overall average of 1.58% and irregular scene text datasets by an overall average of 1.85%. As a result, the proposed approach shows better improvement particularly for recognition of irregular scene text datasets. Therefore, we can conclude that the proposed approach is robust for recognition of irregular scene texts as well as regular scene texts.

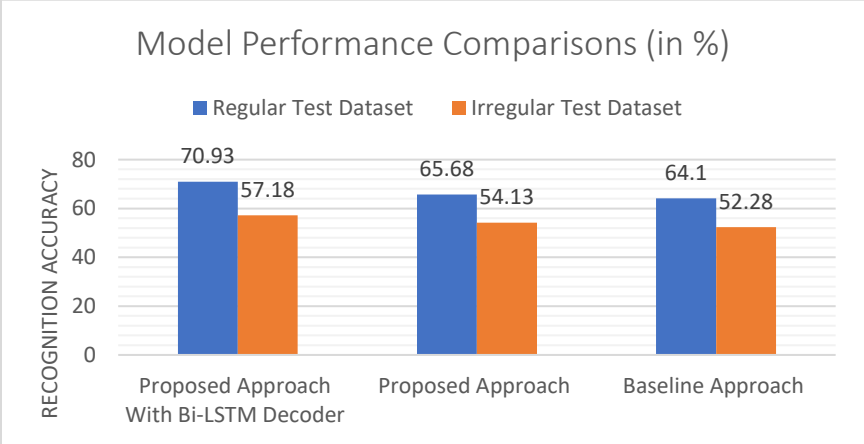


Figure 5.3: Scene Text Recognition Performance Comparison of Proposed Approach, Baseline Approach and Proposed Approach with Bi-LSTM Decoder

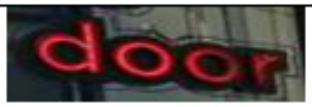
Scene Image	Proposed Approach	Li et al. [19]
	ADDIS ABAL 1899	DSABNG 1899
	ETHIOPIA	ETHIOPIAL
	BALLY	BALAY
	COVIDE19	COVIDES
	TOAST	TOAST
	LONDON	LONDEN
	SOUTHAMPTON	SOUTHAMPIOM
	DISHTA CHALLENGE	DISHTA CHATTENGE
	MERRY	MERRT
	DOOR	DOOR
	CHINESE	CHINESS
	ERMIAS	ERMIIAS

Figure 5.4: Scene Text Recognition Comparison of Proposed Approach and State of The Art Baseline Approach. Green & Red Character Indicates Correctly & Incorrectly Recognized Characters Respectively

5.6 Threats to Validity

5.6.1 Threats to Internal Validity

Internal validity is the degree of confidence that the experiments are not influenced by other internal factors. Threats to internal validity might arise from used measuring or implementing devices, tools and training datasets. Measurement associated threats, specifically threats to instrumentation might affect the validity of experimental results. Those threats might arise from the type of machine used for running the experiments, the type & number of training datasets and the choice of development tools. In this research, however, the proposed approach and state-of-the-art baseline approach are evaluated on similar machine. Both approaches are implemented and trained on similar development tools and similar set of training datasets. For that reason, the threats will not affect the experimental results and comparisons made in this research.

5.6.2 Threats to External Validity

External validity is related to generalization. It is the approximate truth of conclusions that involve generalizations. Threats to external validity might arise from:

- Choice of parameters (strides, input image size, kernel size, sampling type, feature map size, sequence length and others): This threat might affect experiments conducted with proposed approach as well as reproduced baseline work. In this research, both approaches are subjected to similar parameter settings. Hence, threats related to parameter choice will not affect comparisons.
- Selection of evaluation benchmarks: Variety of regular and irregular scene text recognition benchmarks are available. Using them without uniformity for evaluation purpose might also impose another threat to external validity. In this research, however, both approaches are evaluated with similar and ICDAR standard evaluation benchmarks which contains their own uniform set of scene text image datasets.
- Dataset collection methodology: Irrespective of the dataset collection methodology, all training and evaluation datasets should be similar for all experiments. In this research the same datasets are used for both training and testing the proposed and baseline approaches.

Hence, the threats will not affect the experimental results and comparisons made in this research.

Chapter 6

Conclusion and Recommendation

6.1 Conclusion

Different researches are conducted in the area of scene text recognition. Despite a good performance in regular scene text recognition, recognizing irregular and distorted scene texts is a very challenging task. The state-of-the-art approach in scene text recognition is attention-based encoder decoder framework which can handle scene text irregularities implicitly. However, it underperforms since it has a problem of attention mismatch and poor encoder output utilization during each consecutive decoding time steps. To overcome these problems, in this research, we propose an approach based on a global attention mechanism. The proposed approach is evaluated with different regular and irregular scene text dataset classes to assess its recognition capability.

In this research, we analyze the contribution of the global attention mechanism and Bi-LSTM decoder for scene text recognition separately by conducting two experiments. Without any extra supervision information, the proposed global attention mechanism is capable of selecting local image features for decoding characters with LSTM decoder. Being flexible to accommodate different forms of scene text layouts, our approach performs well for both regular and irregular scene texts. Without requiring any explicit image transformation techniques, the proposed approach can handle text distortions implicitly.

Based on the first experiment conducted to see the effect of global attention mechanism on scene text recognition, the proposed approach outperforms the state-of-the-art approach by an average of 1.58% on regular scene text datasets and by an average of 1.85% on irregular scene text datasets. The second experiment is conducted to see the effect of proposed Bi-LSTM decoder over one directional LSTM decoder on scene text recognition, Bi-LSTM decoder contributes an average of 5.24% on recognition of regular scene text datasets and an average of 3.05% on recognition of irregular scene text datasets. From the quantitative analysis of the experimental results, we can conclude that our proposed approach outperforms most of attention based recent works [1], [20],

[23] on both regular and irregular scene text datasets, because of the baseline approach [19] mainly outperforms them.

6.2 Recommendations and Future Works

As a future work, the proposed approach can be extended in different ways. First, it could be integrated with text to speech converter and make the recognized scene texts available for blind peoples and other requiring applications as an audio output. Second, in the feature extraction stage, the ResNet-based convolutional neural network (CNN) can be replaced with latest and robust ResNet-based CNN architecture with a greater number of layers for much better feature extraction.

We recommend to change the proposed system to real-time system by replacing scene text image preprocessing tasks with another real-time and analogues tasks, like integrating it with real-time scene text detectors or any other options. By making this, it is possible to create an end-to-end real-time scene text recognition system.

We also suggest to investigate other options to come up with a more accurate scene text recognition model by applying advanced techniques like transformer-based encoder decoders or other alternatives.

References

[1] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao and X. Bai, "ASTER: An Attentional Scene Text Recognizer with Flexible Rectification," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 1-Sept-2019, Vol. 41(9), pp. 2035-2048.

[2] F. Zhan and S. Lu, "ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2054-2063.

[3] C. Luo, L. Jin and Z. Sun, "MORAN: A Multi-Object Rectified Attention Network for scene text recognition," Pattern Recognition, Vol. 90, 2019, Pages 109-118, ISSN 0031-3203.

[4] M. Yousef, K. F. Hussain and U. S. Mohammed, "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks," Pattern Recognition, Vol. 108, 2020, 107482, ISSN 0031-3203.

[5] Q. Lin, C. Luo, L. Jin and S. Lai, "STAN: A sequential transformation attention-based network for scene text recognition," Pattern Recognition, Volume 111, 2021, 107692, ISSN 0031-3203.

[6] A. A. Gebrie "Ethiopic and Latin Multilingual Text Detection and Script Identification from Videos and Images," M.Sc. thesis in computer engineering from Addis Ababa Institute of Technology, on 10 April 2018, Unpublished.

[7] C. Wang and C.L. Liu, "Multi-branch guided attention network for irregular text recognition," Neurocomputing, Vol. 425, 2021, ISSN 0925-2312, Pages 278-289.

[8] H. Li, D. Yang, S. Huang, K. M. Lam, L. Jin and Z. Zhuang, "Two-dimensional multi-scale perceptive context for scene text recognition," Neurocomputing, Vol. 413, 2020, ISSN 0925-2312, Pages 410-421.

[9] L. Yang, P. Wang, H. Li, Z. Li and Y. Zhang, "A holistic representation guided attention network for scene text recognition," Neurocomputing, Vol. 414, 2020, Pages 67-75, ISSN 0925-2312.

[10] C. Yao, X. Bai, N. Sang, X. Zhou, S. Zhou and Z. Cao, "Scene Text Detection via Holistic, Multi-Channel Prediction," ArXiv preprint ArXiv, Volume abs/1606.09002, 2016: n. pag.

[11] A. Cherian and S. Sebastian, "Automatic localization and recognition of perspectively distorted text in natural scene images," 2016 International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS), 2016, pp. 1-6.

[12] M. Zarechensky, “Text detection in natural scenes with multilingual text”, Department of Analytical Information Systems, Saint Petersburg State University, on 2014.

[13] Y. Zhu, S. Wang, Z. Huang, K. Chen, X. Lin and Q. Zhang, "Attention-Based Text Recognition in Image," 2019 IEEE Fourth International Conference on Data Science in Cyberspace (DSC), 2019, pp. 278-283.

[14] Y. Fang, C. Deyun and W. Rui, "A distortion correction approach on natural scene text image," Proceedings of 2011 6th International Forum on Strategic Technology, 2011, pp. 1058-1061.

[15] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao and X. Bai, "Scene text recognition from two-dimensional perspective," In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8714-8721. 2019.

[16] X. Yang, D. He, Z. Zhou, D. Kifer and C. L. Giles, “Learning to Read Irregular Text with Attention Mechanisms,” Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, {IJCAI-17}, 2017, pp. 3280--3286.

[17] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu and S. Zhou, "Focusing Attention: Towards Accurate Text Recognition in Natural Images," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 5086-5094.

[18] M. Liao, J. Zhang, Z. Wan, F. Xie, J. Liang, P. Lyu, C. Yao and X. Bai, “Scene Text Recognition from Two-Dimensional Perspective,” AAAI, vol. 33, no. 01, pp. 8714-8721, Jul-2019.

[19] H Li, P. Wang, C. Shen and G. Zhang, “Show, Attend and Read: A simple and strong baseline for irregular text recognition,” In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 01, pp. 8610-8617, 2019.

[20] C. Lee and S. Osindero, "Recursive Recurrent Nets with Attention Modeling for OCR in the Wild," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 2231-2239.

[21] Z. Cheng, Y. Xu, F. Bai, Y. Niu, S. Pu and S. Zhou, "AON: Towards Arbitrarily-Oriented Text Recognition," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5571-5579.

[22] W. Liu, C. Chen, K. Y. K. Wong, Z. Su and J. Han, “STAR-NET: A Spatial Attention Residue Network for Scene Text Recognition,” In BMVC, vol. 2, pp. 7, Sep-2016.

[23] B. Shi, X. Wang, P. Lyu, C. Yao and X. Bai, "RARE: Robust Scene Text Recognition with Automatic Rectification," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4168-4176.

[24] J. Zhao, Y. Wang, B. Xiao, C. Shi, J. Jiang and C. Wang, "ASTR: Adversarial Learning Based Attentional Scene Text Recognizer," Pattern Recognition Letters, vol. 138, Oct-2020, pp. 217–222.

[25] S. Sohail, "A Survey of Visual Attention Mechanisms in Deep Learning," In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Dec-2019, pp. 10705-10714.

[26] N. S. Chauhan, "Attention mechanisms in deep learning, explained," arXiv preprint arXiv: 2204.07756, Jan-2021.

[27] J. Lee, S. J. Oh, S. Park, S. Kim, J. Baek and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops (CVPRW), 2020, pp. 2326-2335.

[28] D. Cao, Y. Zhong, L. Wang, Y. He and J. Dang, "Scene text detection in natural images: A Review," Symmetry, vol. 12, no. 12, Nov-2020, pp. 1956.

[29] A. Khan, A. Sohail, U. Zahoora and A. S. Qureshi, "A Survey of the Recent Architectures of Deep Convolutional Neural Networks," Artificial Intelligence Review, April-2020, vol. 53, pp. 5455–5516.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778.

[31] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS), vol. 27, December-2014, pp. 3104–3112.

[32] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho and Y. Bengio, " Attention-based models for speech recognition," Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS), vol. 28, December-2015, pp. 577–585.

[33] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh and H. Lee, "What is wrong with scene text recognition model comparisons? Dataset and model analysis," 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Oct-2019, pp. 4714-4722.

[34] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, “Synthetic data and artificial neural networks for natural scene text recognition,” arXiv preprint arXiv:1406.2227, Jun-2014.

[35] A. Gupta, A. Vedaldi, and A. Zisserman, “Synthetic data for text localisation in natural images,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2315-2324.

[36] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, R. Young, K. Ashida, H. Nagai, M. Okamoto, H. Yamamoto, H. Miyao, J. Zhu, W. Ou, C. Wolf, J. Jolion, L. Todoran, M. Worring and X. Lin. “ICDAR 2003 robust reading competitions: entries, results, and future directions,” International Journal of Document Analysis and Recognition (IJDAR), vol. 7(2), 2005, pp. 105-122.

[37] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan and L. P. de las Heras, “ICDAR 2013 robust reading competition,” 2013 12th International Conference on Document Analysis and Recognition (ICDAR), 2013, pp. 1484-1493.

[38] D. Karatzas, L. G. Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, “ICDAR 2015 competition on robust reading,” 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Aug-2015, pp. 1156-1160.

[39] A. Mishra, K. Alahari, and C.V. Jawahar, “Scene text recognition using higher order language priors,” In BMVC-British machine vision conference. BMVA, Sep-2012.

[40] K. Wang and S. Belongie, “Word spotting in the wild,” In European conference on computer vision (ECCV), pp. 591-604. Springer, Berlin, Heidelberg, 2010.

[41] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, “Recognizing text with perspective distortion in natural scenes,” 2013 IEEE International Conference on Computer Vision (IEEE-ICCV), 2013, pp. 569-576.

[42] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, “A robust arbitrary text detection system for natural scene images,” Expert Systems with Applications, Dec-2014, vol. 41(18), pp. 8027-8048.

[43] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, “ImageNet: A Large-Scale Hierarchical Image Database,” 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June-2009, pp. 248-255.

[44] C. Fellbaum and G. Miller, “WordNet: An Electronic Lexical Database,” in WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 22-22.

[45] A. Graves, S. Fern´andez, F. J. Gomez, and J. Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” In Proceedings of the 23rd international conference on Machine learning (ICML), Jun-2006, pp. 369-376.

[46] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. Zemel and Y. Bengio, “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention,” In International conference on machine learning (ICML), 2015, pp. 2048-2057, PMLR.

[47] W. Liu, C. Chen and K. K. Wong, “Char-Net: A Character-Aware Neural Network for Distorted Scene Text Recognition,” In Proceedings of the 32th AAAI Conference on Artificial Intelligence (AAAI-18), April-2018, vol. 32(1).

[48] R. Shantha, S. Kumari and R. Sangeetha, “Optical character recognition for document and newspaper,” International Journal of Applied Engineering Research, Jan-2015, vol. 10(20), pp. 2015.

[49] G. Vamvakas, B. Gatos, N. Stamatopoulos, and S.J. Perantonis, “A Complete Optical Character Recognition Methodology for Historical Documents,” 2008 The Eighth IAPR International Workshop on Document Analysis Systems, Nov-2008, pp. 525-532.

[50] Q. Ye, Q. Huang, W. Gao and D. Zhao, “Fast and robust text detection in images and video frames,” Image and vision computing, Jun-2005, vol. 23(6), pp.565-576.

[51] F. L. Bookstein, “Principal warps: Thin-plate splines and the decomposition of deformations,” In IEEE Transactions on Pattern Analysis and Machine Intelligence, June-1989, vol. 11(6), pp. 567-585.

[52] M. Jaderberg, K. Simonyan and A. Zisserman, “Spatial transformer networks,” Advances in neural information processing systems (NeurIPS), 2015, vol. 28, pp. 2017–2025.

[53] S. Hochreiter, and J. Schmidhuber, “Long Short-Term Memory,” Advances in neural information processing systems, MIT Press, Presented at NIPS 96 Neural Competition, Nov-1997, vol. 9(8), pp. 1735-1780.

[54] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga and A. Lerer, “Automatic differentiation in PyTorch,” In Proceedings of Advances in Neural Information Processing Systems Autodiff Workshop (NIP- S-W), 2017.

[55] N. Mor, and L. Wolf, “Confidence prediction for lexicon-free OCR,” In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 218-225.

[56] D. Bahdanau, K. Cho and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” In arXiv preprint arXiv: 1409.0473, Sep-2014, vol. 1409.0473.