Journal of Autonomous Intelligence (2024) Volume 7 Issue 5
doi: 10.32629/jai.v7i5.1334

ORIGINAL RESEARCH ARTICLE

STRTrans: An accurate scene text recognition based on improved transformer network

Prabu Selvam[1], Saravanan Palani[1,*], Marimuthu M[1,*], Elakkiya Rajasekar[2]

[1] *School of Computer Science and Engineering, Vellore Institute of Technology, Chennai 600127, India*

[2] *Department of Computer Science and Information System, BITS Pilani, Dubai 345055, UAE*

*** Corresponding authors:** Saravanan Palani, saravanan.p@vit.ac.in; Marimuthu M, marimuthu.m@vit.ac.in

ABSTRACT

Text recognition represents a significant research domain within the field of computer vision. Specifically, scene text recognition (STR), which involves the identification of text within real-world scenes, presents a distinctive set of challenges. These challenges encompass the need for text to capture attention immediately, the potential for text distortion, and the influence of various factors like occlusion, noise, and obstructions during the image capture process. All of these elements significantly complicate the task of recognizing text within scenes. In this paper, we introduce STRTrans, a modified Transformer network designed to enhance the performance of STR. This enhancement addresses the shortcomings observed in the existing model, characterized by lower accuracy and difficulties in recognizing irregular text. The modification of the encoder structure involves the implementation of two consecutive layers of the self-attention (SA) mechanism and the reduction of the point-wise feed-forward layer. This modification aims to enable the network to interpret the semantic arrangement better. Our approach underwent experimental validation using three publicly available datasets and was benchmarked against other advanced methods. The experimental results consistently demonstrate the robust performance of our approach across all three benchmark tests, achieving recognition accuracies of 90.60%, 86.20%, and 86.90% in the IC15, SVT-P, and CUTE datasets, respectively. Moreover, the improved model comprehensively surpasses the existing approaches.

Keywords: text recognition; deep learning; transformer; attention; image rectification

ARTICLE INFO

Received: 10 October 2023
Accepted: 21 November 2023
Available online: 29 July 2024

1. Introduction

Textual information holds significant value in computer vision-based applications such as product retrieval, key information extraction, ID card scans, autonomous driving, and travel translators. Extracting textual information is a critical and foundational aspect of scene text recognition. STR's sequence-to-sequence (S2S) approach offers superior results in recognizing grayscale fonts. Due to technological advancements, scene text presents itself in various font styles, colours, and layouts, with blur, noise, and complex backgrounds. Consequently, given the current landscape, extracting Scene Text (ST) has become complex[1]. ST can be categorized as either regular or irregular. In the case of regular ST, the character sequence is extracted directly from straight-text images, making the process straightforward and categorized as image-based sequence recognition. However, this approach does not yield satisfactory results with complex and erratic fonts and varying styles. Hence, attention-based and connectionist temporal classification techniques are preferred for extracting complex ST. On the other hand, Irregular

ST possesses an unconventional typeface with unclear and curved shapes, making identification challenging. Current methods for irregular ST recognition involve rectification, multi-direction encoding, and attention-based algorithms[2].

Convolutional Neural Networks (CNNs) can capture spatial hierarchies of features, making them excellent at pattern and image recognition[3]. They employ parameter sharing and spatial invariance to identify patterns in any location within an image. During training, CNNs automatically pick up on pertinent features with the help of methods like data augmentation and regularization. They are a strong option for image identification jobs due to their high scalability and efficiency benefits from pre-trained models and parallelization. A CNN can be employed for ST extraction, wherein individual characters are recognized and detected in a traditional CNN paradigm. However, the performance of the CNN tends to degrade when characters are intertwined or embedded within others. As a result, contemporary approaches favour end-to-end ST recognition, as the accuracy of detection hinges on the character detector.

A Recurrent Neural Network (RNN) is integrated with the CNN model to enhance accuracy and recognition capabilities. RNNs excel in capturing long-term dependencies in context, but they are not suited for parallel processing and require more time for sequential information analysis. The classification of irregular text recognition encompasses multi-direction encoding, rectification, character-level intensive supervision, data-guided approaches, and 2D perspective-based recognition. Conventional data augmentation strategies fail to achieve high accuracy due to the significant distortions present in irregular ST. Consequently, including multiple training samples is necessary to account for these distortions effectively[4].

The remainder of this article is organized as follows: Section 2 delves into existing works related to ST extraction and recognition, along with their respective advantages and disadvantages. Section 3 discusses the requisite system architecture and equations, detailing the proposed model. Section 4 presents the experimental results of the suggested method and includes a comparative analysis with existing works. Finally, Section 5 concludes the proposed work and outlines potential avenues for future improvements.

2. Related works

In this section, we delve into the existing work related to scene text extraction and recognition, exploring their merits and limitations.

Zhang et al.[5] introduced the scale-aware hierarchical attention network (SaHAN) for STR. This approach leverages the pyramidal structure inherent to deep CNNs, maintaining multi-scale properties for adjustable receptive fields. The hierarchical attention decoder is applied twice to multi-scale characteristics, providing the most granular data for prediction. The SaHAN approach requires only images and their corresponding training text labels. While it effectively addresses character scale-variation issues, it does not tackle other challenges. Mu et al.[6] proposed random blur data augmentation for STR. This procedure employs Random Blur Regions (RBR) and Random Blur Units (RBU). RBR generates potentially confusing samples during training, leading to lower recognition accuracy due to ambiguous training samples. To mitigate this issue, the number of subunits is divided into RBUs, which enhances model training and improves sample readability. RBUs perform exceptionally well when there is sufficient, but not excessive, training data.

Qiao et al.[7] introduced the parallel, iterative, and mimicking network (PIMNet) for STR. PIMNet balances accuracy and efficiency, achieving faster text prediction through parallel attention mechanisms compared to sequential attention. Furthermore, iterative prediction enhances prediction accuracy. In imitation learning, this approach employs two decoders—analogous to a parallel decoder and an autoregressive decoder—to enhance hidden layer performance. PIMNet supports the complete training cycle

without prerequisite training for the end-to-end process, outperforming more complex techniques within a typical framework.

Phan et al.[8] proposed extracting perspective-distorted text in natural scenes using the bag of keywords. The pre-training phase involves the use of scale-invariant feature transform descriptors. Lexicon terms use word recognizers, but support for arbitrary orientation is limited. Liu et al.[9] employed the spatial attention residue network (STAR-Net) for STR. This approach corrects distortions in natural images using a spatial transformer, enhancing feature extraction while maintaining minimal distortion. Residue convolutional blocks extract the text's discriminative properties, facilitating fine-grained recognition. Combining STAR-Net and residue convolutional blocks creates an end-to-end trainable network for effective STR. Selvam et al.[10] discussed deep learning strategies for detecting and identifying supermarket products, including retail product detection, product text detection, and product text recognition. They utilized the YOLOv5 object detection algorithm, ResNet50, and FPN to increase text detection and recognition accuracy for regular and irregular text scenarios.

Lee et al.[11] employed recursive recurrent nets with attention modeling for OCR in the wild. This approach involves image feature extraction using recursive CNNs and incorporates RNNs in character-level language models to avoid using n-grams. The end-to-end training procedure substitutes the soft-orientation approach for conventional back-propagation, benefiting both constrained and unconstrained scenarios. Risnumawan et al.[12] proposed text detection for natural scene images, utilizing properties such as mutual direction symmetry, magnitude symmetry, and gradient vector symmetry to identify text pixel possibilities. The SIFT method is used for pixel-perfect text. The nearest neighbour criteria are employed during the ellipse-growing procedure to extract text components while separating non-text elements based on text direction and spatial analysis. Yu et al.[13]harnessed Semantic Reasoning Networks (SRN) for accurate STR, considering both semantic data and visual textures. SRN is introduced to address the limitations of RNNs, with the Global Semantic Resonating Module (GSRM) collecting global semantic context. This technique effectively recognizes long texts in regular, irregular, and non-Latin languages.

Xia et al.[14] proposed a STR approach based on a two-stage attention and multi-branch feature fusion module. A two-stage attention technique reminiscent of a transformer-based encoder-decoder structure is used to capture text in STR effectively. This method involves initially extracting text from an image and then determining its location, thereby increasing prediction accuracy. The multi-branch feature function is employed to enhance accuracy by incorporating more features. Wu et al.[15] introduced STR's two-level rectification attention network (TRAN) involving text rectification and recognition levels. The attention recognition network identifies texts in rectified images after correction at the pixel and geometry levels by the two-level rectification network. The channel and kernel-wise attention units are applied to improve feature extraction accuracy. This work adopts early stop training to ensure a smooth convergence process. Dai et al.[16] employed a scale-adaptive orientation attention network for STR, incorporating a sequence recognition network for character-level receptive attention. The dynamic log-polar Transformer is used to learn the log-polar origin for performing arbitrary rotations.

Luan et al.[17] introduced a streamlined transformer network to mitigate attention drift and reduce computational overhead. Their modifications to the Vision Transformer network included the integration of a positional-enhancement block, dynamically fusing positional information with visual data, ultimately leading to enhanced recognition accuracy. In contrast, Selvam et al.[18] employed a standard transformer network to identify regular and irregular text within low-resolution word images. Their approach involved the introduction of an improved SA mechanism known as the threshold-based attention mechanism. This mechanism effectively eliminated less significant elements from the attention matrix. While these methods demonstrated improved recognition outcomes for low-resolution images, their efficacy in handling intricate or embellished word images remained limited.

Kwon et al.[19] introduced the ensemble method "textfooler" for conducting a black-box attack on unfamiliar models generated by the ensemble adversarial. This method replaces specific crucial keywords, words, or phrases to manipulate their meanings. The WordCNN, WordLSTM, and BERT models were integrated into the approach to enhance its effectiveness and success rate[20]. However, achieving a 100% success rate remains challenging, and the method does not accommodate heterogeneous architectures.

Xue et al.[21] proposed the "Image-to-Character-to-Word" technique for STR. This method involves two interconnected tasks: image-to-character, which relies on visual features to detect characters, and character-to-word, which decodes characters to identify text. Unlike the conventional encoder-decoder architecture, this approach directly learns from the image data, reducing the impact of image noise. However, it faces limitations in recognizing scene text due to occlusions and ultra-low resolution.

Vision transformers play a crucial role in enhancing the performance of image-based tasks. Yan et al.[22] proposed an adaptive n-gram transformer for multi-scale STR, automatically selecting the image patch as a significant component for extracting features from multi-scale scene texts. The existing visual model focuses solely on text characters without incorporating linguistic information, resulting in a lower model recognition success rate. Yang et al. [23] introduced the display-semantic Transformer to address this issue. This model can extract semantic information from images, facilitating STR.

From the literature review, previous research has demonstrated that the existing approaches have achieved more remarkable performance in recognizing regular scene text. However, these techniques encounter difficulties in recognizing irregular scene text. The existing approaches overlook the usage of text recognition modules, leading to performance degradation. Additionally, these techniques perform well on high-resolution word images, but their performance is unsatisfactory on low-resolution word images.

3. The proposed model

The Transformer concept, introduced by Vaswani et al.[24], emerged as a remedy for addressing the limitations inherent in Recurrent Neural Networks (RNNs) and encoder-decoder architectures. Their pivotal contribution involved a substantial redesign of the architecture. They achieved this by substituting RNNs with attention mechanisms within the Sequence-to-Sequence (S2S) encoder-decoder framework[18]. Incorporating attention mechanisms empowers the model to uphold long-term memory, meticulously attending to every token created throughout the entire sequence's history. This architectural framework consists of an amalgamation of feed-forward layers (FFL), normalization layers, and residual connections, all thoughtfully stacked on top of each other. Several multi-head attention layers further complement these elements. For a visual representation, please refer to **Figure 1**, which illustrates the proposed STR network.

3.1. Image transformation using TPS++

TPS++[25] comprises two pivotal components: Multi-scale Feature Aggregation (MSFA) and Attention-Incorporated Parameter Estimation (AIPE). Distinguished from existing rectification methods, TPS++ boasts unique attributes. Firstly, it optimizes sharing the visual feature extractor, fostering a seamlessly integrated framework. This innovation effectively manages parameters and speeds up inference while preserving its adaptability. Secondly, it introduces an attention mechanism to TPS, amplifying its capacity for adaptable, content-aware corrections. These advancements collectively elevate rectification quality and streamline recognition. Moreover, TPS++ inherits vital qualities, being end-to-end trainable with STR, eliminating the need for supplemental annotations beyond text labels.

3.1.1. Multi-Scale Feature Aggregation (MSFA)

Prior text rectification methods, as exemplified by Shi et al.[26] and Shi et al.[27], typically served as a preliminary step in STR. However, they imposed a significant computational burden due to the need for

dual-feature extraction. TPS++ addresses this challenge by integrating its feature backbone with the recognizer. A key component, MSFA, processes feature maps from the initial three backbone blocks, resizing and combining them.

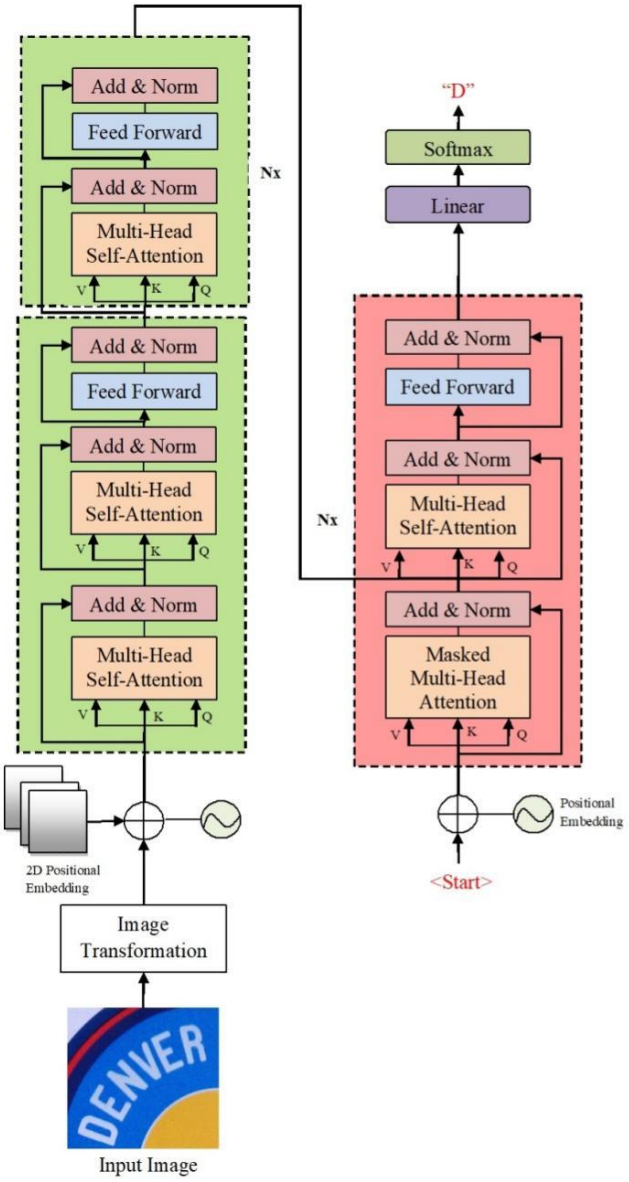

Figure 1. Pipeline of the proposed STRTrans network.

The first and second block features are downsized by 4×, they all maintain 64 standardized feature channels. The encoder-decoder feature extractor employs a streamlined approach, featuring both contracting and symmetric expansive paths with three convolution layers[28].

Additionally, channel spatial joint attention enhances crucial features. These operations fine-tune multi-scale visual features for rectification. Encoded (F_e) and decoded (F_d) features maintain matching channel counts, with (F_d) preserving the scaled input feature's spatial resolution. This division of feature extraction into two components is instrumental in TPS++'s effectiveness. The first emphasizes general visual features, while the second, MSFA, focuses on rectification optimization by leveraging location-related cues from shallower blocks, facilitating control points regression and attention modelling. Notably, the second component introduces minimal computational overhead.

3.1.2. Attention-Incorporated Parameter Estimation (AIPE)

AIPE, positioned after MSFA, enables control point regression and precise content-based attention score estimation. This process is accomplished through a gated attention mechanism, which predicts attention scores tailored to textual content.

In a departure from the traditional method of placing control points along image borders, AIPE uniformly disperses them across the feature map in a grid-like pattern. This strategy optimizes control point distribution, concentrating them within the textual foreground, thus enhancing their relevance while reducing their presence in less important border regions. The number of control points matches the spatial resolution of (F_e) and is then transformed into a feature sequence. Two linear layers are applied to predict x and y offsets for each control point, successfully regressing control points.

In the attention score estimation process, AIPE analyses the relationship between (F_e) and (F_d) using the dynamic gated-attention block. Here, attention scores dynamically capture the interplay between control points and text. The resulting feature undergoes reshaping and is combined with (F_e) through matrix multiplication. Scaling with a factor of $1/\sqrt{D}$ and a Tanh activation function constrain attention scores within the range of (−1, 1).

3.2. Positional Encoding (PE)

Language modelling sequences consist of a fixed token order. While RNNs automatically encode token positions during operation, attention mechanisms do not consider word placement. In contrast, attention-based models can handle encoded words without adherence to order, potentially introducing randomization. Unlike recurrent networks, multi-head attention networks don't naturally leverage word order in input sequences[29]. An approach to address this is encoding each word based on its position in the current sequence. After embedding each word using a matrix, PE captures word positions using the following Equations (1) and (2):

$$PE(p_w, 2 * p_e) = sin\left(\frac{p_w}{\frac{10000^{2*p}e}{d_{model}}}\right) \quad (1)$$

$$PE(p_w, 2 * p_e) = cos\left(\frac{p_w}{\frac{10000^{2*p}e}{d_{model}}}\right) \quad (2)$$

The PE matrix incorporates the sine variable in even positions and the cosine variable in odd positions. d_{model} represents the embedding dimension, p_w indicates the position within the sequence (ranging from 0 to n-1), and p_e signifies the position within the embedding dimension (ranging from 0 to d_{model}).

The position encoding employs a linear transformation layer and a normalization layer. These layers enable the model to leverage positional information effectively, enhancing text recognition accuracy. Additionally, the design of the position branch integrates the concept of a residual connection, facilitating improved utilization of preceding information and further enhancing the branch's performance.

3.3. Multi-Head Self-Attention Mechanism (MHSA)

The MHSA layer consists of numerous attention heads, each computing attention across its input elements: Value (V), Key (K) and Query (Q), subjecting them to linear transformations. This process equips the model to focus simultaneously on diverse representational sub-spaces, fostering richer representations than a single-pass attention mechanism. Unique linear transformations are applied to V, K, and Q components for each attention head, promoting the learning of diverse representations. With N, parallel attention layers,

heads, queries, keys, and values are projected via separate dense layers with hidden sizes q, k, and v. Subsequently, another dense layer processes the concatenated results of these N heads. **Figure 2** depicts the pipeline of the MHSA mechanism. The SA head output follows Equations (3)–(8) for the l[st] Transformer layer.

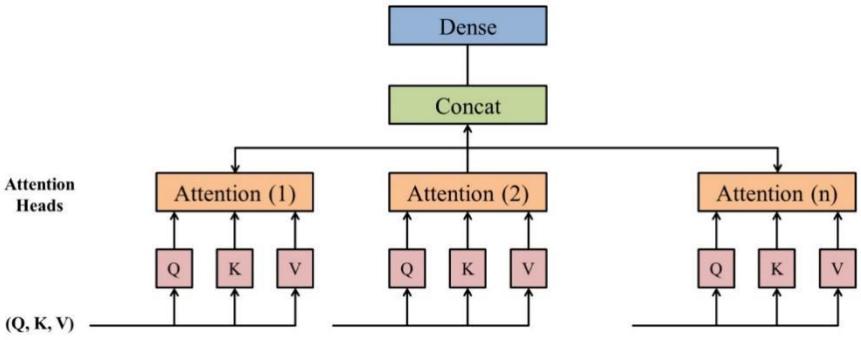

Figure 2. Multi Head Self-Attention Mechanism.

$$Q = [q_1, q_2, \dots, q_w]^T, \qquad q_i = W_q x_i + b_q \tag{3}$$

$$K = [k_1, k_2, \dots, k_w]^T, \qquad k_i = W_k x_i + b_k \tag{4}$$

$$V = [v_1, v_2, \dots, v_w]^T, \qquad v_i = W_v x_i + b_v \tag{5}$$

$$Attention(Q, K, V) = softmax\left(\frac{Q \times K^T}{\sqrt{d_k}}\right)V \tag{6}$$

$$MHAttention(Q, K, V) = Concat(head_1, \dots, head_n)W^O \tag{7}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{8}$$

The variables b, W_q, W_k, and W_v correspond to the bias and weight matrices related to query, key, and value, respectively. The scaling factor $\frac{1}{\sqrt{d_k}}$, which reflects the dimensions of queries and keys, is employed to prevent a minimal gradient within the softmax function.

3.4. The modified encoder block

In the standard Transformer design, as depicted in **Figure 3**, each encoder block consists of a SA mechanism and a FFL, performing diverse roles, including residual calculation and normalization. The FFL functions as a densely connected middle layer, extracting essential information from the word image and fostering active learning among neurons. However, our investigation yielded a significant insight. When we introduced the FFL for non-linear conversion early in the training process, it posed challenges for the text recognition model in capturing authentic, hidden representations within the corpus.

The root of this issue lies in the FFL's use of the Relu activation function, which led to the conversion of all negative values to 0. While the authentic intent was to facilitate system convergence, it inadvertently impeded progress during the initial training phases. To address this concern, we conducted various experiments exploring fusions of SA mechanisms and FFL within the encoder. Our objective was to identify combinations that would not hinder the text recognition model's training progress while facilitating convergence. Our findings revealed that three SA mechanisms are adequate. Furthermore, introducing the SA mechanism earlier in the model, with a more significant stacking than the number of FFL, significantly reduced the language model's perplexity.

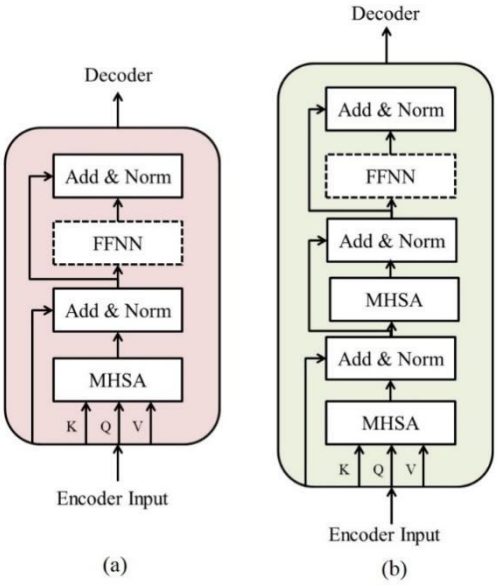

Figure 3. (a) Encoder block of the conventional Transformer; **(b)** Encoder block of the modified Transformer.

Table 1 illustrates the diverse combinations of SA mechanisms and FFL that regulate the model. The symbol "s" signifies an SA mechanism within the table, while "f" denotes an FFL. The sequence of these symbols reflects the specific order in which the layers are arranged. Consequently, the notation "sa-ff+sa-ff+sa-ff+sa-ff+sa-ff+sa-ff" represents the fundamental configuration of the conventional Transformer network, serving as the baseline for comparison.

Table 1. Different configurations of FFL and SA mechanisms in the modified encoder block.

Modified encoder block	Avg. recognition score (%)
sa-ff+sa-ff+sa-ff+sa-ff+sa-ff+sa-ff (Baseline)	70.22
sa+sa-ff+sa+sa-ff+sa+sa-ff	76.54
sa+sa-ff+sa-ff (STRTrans)	87.92
sa+sa-ff+sa-ff+sa-ff	82.36
sa+sa-ff+sa+sa-ff	80.45

Each layer of the Transformer network is comprised of a SA sublayer, followed by a feed-forward sublayer. These sublayers work to modify a sequence of vectors denoted as M_0, using Equations (9)–(13):

$$M_1 = Self - Attention\ (M_0) + M_0 \quad (9)$$

$$M_2 = Self - Attention\ (M_1) + M_1 \quad (10)$$

$$M_3 = FFN\ (M_2) + M_2 \quad (11)$$

$$M_4 = Self - Attention\ (M_3) + M_3 \quad (12)$$

$$M_5 = FFN\ (M_4) + M_4 \quad (13)$$

After a series of experiments, we discovered that three SA mechanisms proved to be adequate. It appears that the encoder encounters the learning plateau phenomenon once more. Ultimately, we opted for the "sa+sa-ff+sa-ff" as our proposed model due to its lower count of model parameters. We also discovered that if the SA mechanism is stacked beyond the number of FFL, the language model's perplexity can significantly decrease.

3.4. Decoder block

The decoder inputs query, key, and value parameters from the initial SA mechanism, following the sub-word embedding operation and the ground truth. A masked MHSA was incorporated to halt the network from anticipating the valid, future training sequence.

The masked MHSA was applied by multiplying it with a negative infinity value, assuming that the input sequence of the decoder follows the pattern: $d, d+1, d+2, \ldots, d+n$. During the training step t, every sequence other than the input sequence d, such as $d+1, d+2, \ldots, d+n$, was also multiplied by the negative infinity value. This process was repeated during the subsequent training steps, such as $t+1$, where the sequence $d+2, \ldots, d+n$, excluding d and $d+1$, was similarly affected. This iteration continued until the final training step, $t+n$, marking the conclusion of the mask MHSA implementation.

The decoder's subsequent SA mechanism computed the correlation between the output from the encoder and the output from the primary SA mechanism in the decoder. The parameter Q was used to aggregate and standardize the input of the first SA mechanism in the decoder and the residual outcome from the first SA mechanism. Furthermore, the K and V parameters represent the encoder's output.

The "Encoder-Decoder Attention" layer in the decoder section functions similarly to the MHSA mechanism. However, it generates its Q matrix from the underlying layer while utilizing the K and V matrix from the output of the encoder stack. The Linear layer, a primary fully connected neural network, projects the vector derived from the stack of decoders into a significantly larger vector known as a logits vector. Subsequently, the softmax layer transforms these scores into probabilities, all of which are positive and collectively sum up to 1.0. The cell with the highest probability is selected, and the associated word is then generated as the output for the current time step.

4. Experimental results and discussion

This section provides a concise and precise description of the experimental results, their interpretation, as well as the experimental conclusions that can be drawn.

4.1. Dataset details

The STRTrans network was trained using two synthetic datasets, MJSynth[30] and SynthText[8]. Furthermore, the network underwent training and testing on three benchmark datasets, namely ICDAR2015 (IC15)[31], SVT-Perspective (SVT-P)[32], and CUTE80[33]. **Table 2** provides descriptions of three benchmark datasets.

Table 2. Dataset Description.

Dataset	Total images	Training images	Testing images	Image characteristics
IC15	6,545	4,468	2,077	Irregular shapes (horizontal, oriented, curved); Captured by Google Glasses without precise positioning.
SVP-P	645	516	129	Heavily distorted, noisy, blurred, low-resolution images; Captured from a side-view angle using Google Street View.
CUTE80	80	64	16	Arbitrarily shaped letters; High-resolution images captured in naturalistic settings; Evaluates irregular STR.

4.2. Implementation details

In our SVTR approach, we utilize Zheng et al.'s[25] rectification module to standardize image text to 32 × 64 dimensions, thereby correcting distortions. Our training employs the Adam optimizer with a weight decay of 0.05 to enhance model performance. We set an initial learning rate of 0.01 with a batch size of 16 to facilitate convergence. We employ data augmentation techniques to enhance model robustness, including

random rotations, perspective distortions, motion blurs, and Gaussian noise. Our alphabet encompasses all case-insensitive alphanumeric characters, and we limit the maximum prediction length to 25 characters. Word accuracy serves as the primary evaluation metric, and we use Tesla T4 GPUs on Google Colaboratory for efficient training within the PyTorch framework. **Table 3** presents a summary of the parameters for the STRTrans.

Table 3. STRTrans model's parameters.

Parameter	STRTrans
Optimizer	Adam
Learning rate	0.01
# of Epochs	200
Activation function	Softmax
Batch size	16
Dataset split ratio (Training: Validation: Testing)	80:10:10
Weight decay	0.05

4.3. Results

Figure 4 depicts the accuracy graph of the proposed approach, which reaches 90.6% on the IC15 training dataset at the 40th epoch, with no further improvement observed until the 100th epoch. Additionally, it achieves an accuracy of 91.3% on the IC15 test dataset. In **Figure 5**, the accuracy graph of the proposed approach is shown. The corresponding loss valuers are 0.25 and 0.14 on the taining and validation IC15 datasets.

Figure 6 illustrates the sample results of the proposed methods on three benchmark datasets. The first image shows the input image, the second image shows the attention map and the final image shows the predicted text.

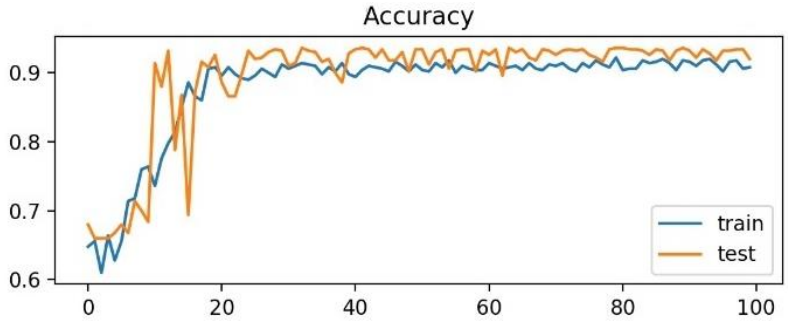

Figure 4. Accuracy graph on IC15 dataset.

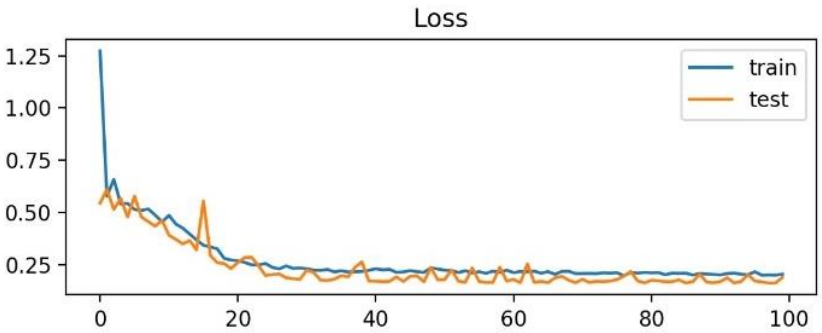

Figure 5. Loss graph on IC5 dataset.

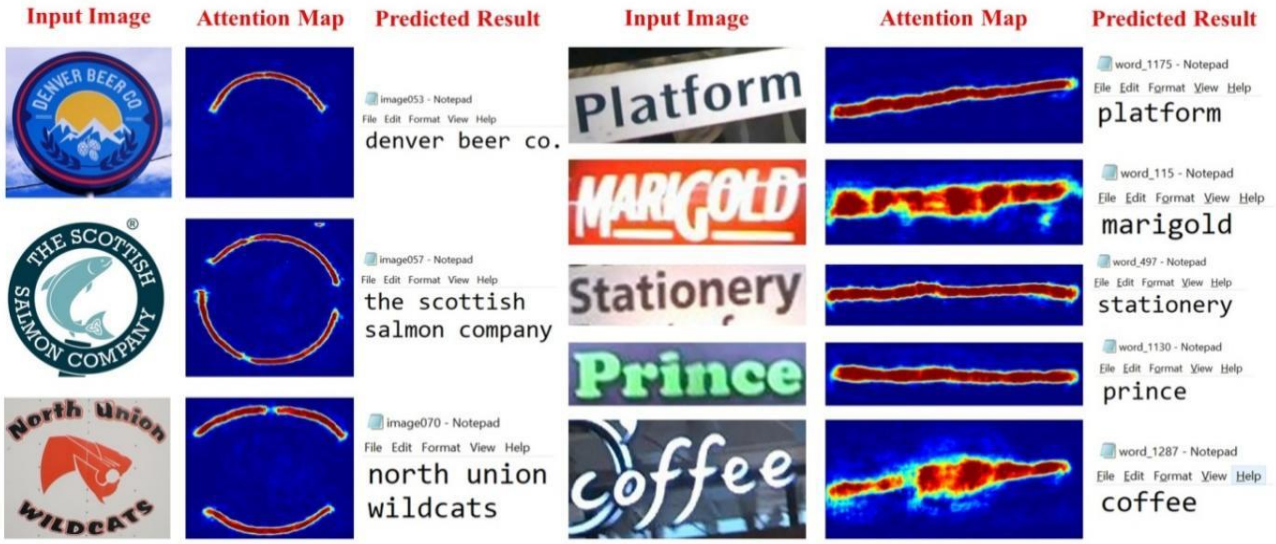

Figure 6. Visualization of a sample experimental result.

4.4. Performance comparison

The performance comparison between the STRTrans network and various existing approaches in scene text recognition highlights the remarkable capabilities of the STRTrans network, as shown in **Table 4**. Existing methods like RARE[26], Rosetta[33], STAR-Net[8], SEED[33], R2AM[11], and Text is Text[29] exhibit respectable accuracy percentages on datasets like IC15, SVT-P, and CUTE. However, they fall short in recognition accuracy compared to the proposed method. For instance, RARE[26] achieves accuracy percentages ranging from 70.40% to 76.20%, while Rosetta[33], STAR-Net[8], and SEED[33] demonstrate competitive but lower accuracy. R2AM[11] and Text is Text[29] record good results but are surpassed by the proposed method.

Table 4. Performance comparison with existing approaches.

Methods	**Irregular datasets (Accuracy %)**		
	IC15	**SVT-P**	**CUTE**
RARE[26]	74.50	76.20	70.40
Rosetta[33]	71.20	73.80	69.20
STAR-Net[8]	76.10	77.50	71.70
SEED[33]	80.00	81.40	83.60
R2AM[11]	68.90	72.10	64.90
Text is Text[29]	76.90	84.40	86.30
Selvam et al.[18]	88.20	**90.60**	**91.30**
STRTrans	**90.60**	**86.20**	**86.90**

In contrast, the STRTrans consistently outperforms these existing approaches, achieving exceptional recognition accuracy of 90.60% on IC15, 86.20% on SVT-P, and 86.90% on CUTE. Selvam et al.[18] implemented bi-directional embedding in the decoder section, resulting in an enhanced performance with an accuracy of 88.20% on IC15, 90.60% on SVT-P, and 91.30% on CUTE. In contrast, the proposed STRTrans outperforms Selvam et al.[18] in the IC15 dataset with an accuracy of 90.60% but also achieves the second-highest accuracy of 86.20% on SVT-P and 86.90% on CUTE.

This superior performance establishes the STRTrans network as a prominent choice for STR, striking a remarkable balance between accuracy and practicality. While each existing method has its merits, such as

addressing specific text recognition challenges, the proposed approach's exceptional recognition accuracy solidifies its position as a leading solution in the field.

5. Conclusion

In our research, we have introduced STRTrans, an enhanced Transformer network specifically designed to push the boundaries of STR to new heights. Our crucial innovation revolves around streamlining the point-wise FFL operation—a critical step in bolstering the model's grasp of semantic information. Our model has significantly enhanced its ability to understand input sequences and the corresponding semantic nuances during training by introducing supplementary SA mechanisms before this layer. We have subjected the STRTrans network to rigorous evaluation using three widely recognized public datasets. The results have revealed substantial advancements in accuracy when compared to the original Transformer network. Additionally, we have conducted comprehensive comparative analyses with other experimental methods, further affirming the robust performance of our improved network. This research underscores the immense potential of STRTrans in advancing the field of STR. It promises to enable more precise and efficient textual content recognition across diverse real-world scenarios, marking a significant step forward in this vital study area. Our forthcoming research aims to expand the proposed method into a comprehensive end-to-end STR pipeline. Additionally, we intend to investigate and mitigate potential adversarial attacks on text as part of our research agenda.

Author contributions

Conceptualization, PS, SP and ER; methodology, PS; software, MM; validation, PS, SP and MM; formal analysis, ER; investigation, SP; resources, PS; data curation, SP, MM and ER; writing—original draft preparation, PS; writing—review and editing, SP; visualization, MM; supervision, SP, MM and ER; project administration, SP and MM. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Yang M, Yang B, Liao M, Zhu Y. and Bai X. Class-Aware Mask-guided feature refinement for scene text recognition. Pattern Recognition, 2024, 149: 110244. doi: 10.1016/j.patcog.2023.110244
2. Lu N, Yu W, Qi X, et al. MASTER: Multi-aspect non-local network for scene text recognition. Pattern Recognition. 2021, 117: 107980. doi: 10.1016/j.patcog.2021.107980
3. Sengan S, Priya V, Syed Musthafa A, et al. A fuzzy based high-resolution multi-view deep CNN for breast cancer diagnosis through SVM classifier on visual analysis. Varadarajan V, Kommers P, Piuri V, Subramaniyaswamy V, eds. Journal of Intelligent & Fuzzy Systems. 2020, 39(6): 8573-8586. doi: 10.3233/jifs-189174
4. Wang C, Liu CL. Multi-branch guided attention network for irregular text recognition. Neurocomputing. 2021, 425: 278-289. doi: 10.1016/j.neucom.2020.04.129
5. Zhang J, Luo C, Jin L, et al. SaHAN: Scale-aware hierarchical attention network for scene text recognition. Pattern Recognition Letters. 2020, 136: 205-211. doi: 10.1016/j.patrec.2020.06.009
6. Mu D, Sun W, Xu G, et al. Random Blur Data Augmentation for Scene Text Recognition. IEEE Access. 2021, 9: 136636-136646. doi: 10.1109/access.2021.3117035
7. Qiao Z, Zhou Y, Wei J, et al. PIMNet: A Parallel, Iterative and Mimicking Network for Scene Text Recognition. Proceedings of the 29th ACM International Conference on Multimedia. Published online October 17, 2021: 1-10. doi: 10.1145/3474085.3475238
8. Phan TQ, Shivakumara P, Tian S, et al. Recognizing Text with Perspective Distortion in Natural Scenes. 2013 IEEE International Conference on Computer Vision. Published online December 2013: 1-13. doi: 10.1109/iccv.2013.76
9. Liu W, Chen C, Wong KYeeK, et al. STAR-Net: A SpaTial Attention Residue Network for Scene Text Recognition. Procedings of the British Machine Vision Conference 2016. Published online 2016. doi: 10.5244/c.30.43

10. Selvam P, Koilraj JAS. A Deep Learning Framework for Grocery Product Detection and Recognition. Food Analytical Methods. 2022, 15(12): 3498-3522. doi: 10.1007/s12161-022-02384-2
11. Lee CY, Osindero S. Recursive Recurrent Nets with Attention Modeling for OCR in the Wild. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Published online June 2016: 2231-2239. doi: 10.1109/cvpr.2016.245
12. Risnumawan A, Shivakumara P, Chan CS, et al. A robust arbitrary text detection system for natural scene images. Expert Systems with Applications. 2014, 41(18): 8027-8048. doi: 10.1016/j.eswa.2014.07.008
13. Yu D, Li X, Zhang C, et al. Towards accurate scene text recognition with semantic reasoning networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Published online June 2020: 1-10. doi: 10.1109/cvpr42600.2020.01213.
14. Xia S, Kou J, Liu N, et al. Scene text recognition based on two-stage attention and multi-branch feature fusion module. Applied Intelligence. 2022, 53(11): 14219-14232. doi: 10.1007/s10489-022-04241-5
15. Wu L, Xu Y, Hou J, et al. A Two-Level Rectification Attention Network for Scene Text Recognition. IEEE Transactions on Multimedia. 2023, 25: 2404-2414. doi: 10.1109/tmm.2022.3146779
16. Dai P, Zhang H, Cao X. SLOAN: Scale-Adaptive Orientation Attention Network for Scene Text Recognition. IEEE Transactions on Image Processing. 2021, 30: 1687-1701. doi: 10.1109/tip.2020.3045602
17. Luan X, Zhang J, Xu M, et al. Lightweight Scene Text Recognition Based on Transformer. Sensors. 2023, 23(9): 4490. doi: 10.3390/s23094490
18. Selvam P, Koilraj JAS, Romero CAT, et al. A Transformer-Based Framework for Scene Text Recognition. IEEE Access. 2022, 10: 100895-100910. doi: 10.1109/access.2022.3207469
19. Kwon H, Lee S. Detecting textual adversarial examples through text modification on text classification systems. Applied Intelligence. 2023, 53(16): 19161-19185. doi: 10.1007/s10489-022-03313-w
20. Kwon H, Lee S. Ensemble transfer attack targeting text classification systems. Computers & Security. 2022, 117: 102695. doi: 10.1016/j.cose.2022.102695
21. Xue C, Huang J, Zhang W, et al. Image-to-Character-to-Word Transformers for Accurate Scene Text Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. Published online 2023: 1-14. doi: 10.1109/tpami.2022.3230962
22. Yan X, Fang Z, Jin Y. An adaptive n-gram transformer for multi-scale scene text recognition. Knowledge-Based Systems. 2023, 280: 110964. doi: 10.1016/j.knosys.2023.110964
23. Yang X, Silamu W, Xu M, et al. Display-Semantic Transformer for Scene Text Recognition. Sensors. 2023, 23(19): 8159. doi: 10.3390/s23198159
24. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of the 31st on Neural Information Processing Systems (NIPS 2017); Long Beach, CA, USA, 2017: 1-11.
25. Zheng T, Chen Z, Bai J, et al. TPS++: Attention-Enhanced Thin-Plate Spline for Scene Text Recognition. Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence. August 2023: 1777-1785. doi: 10.24963/ijcai.2023/197
26. Shi B, Wang X, Lyu P, et al. Robust Scene Text Recognition with Automatic Rectification. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Published online June 2016: 4168-4176. doi: 10.1109/cvpr.2016.452
27. Shi B, Yang M, Wang X, et al. ASTER: An Attentional Scene Text Recognizer with Flexible Rectification. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2019, 41(9): 2035-2048. doi: 10.1109/tpami.2018.2848939
28. Prabu S, Joseph Abraham Sundar K. Enhanced Attention-Based Encoder-Decoder Framework for Text Recognition. Intelligent Automation & Soft Computing. 2023, 35(2): 2071-2086. doi: 10.32604/iasc.2023.029105
29. Bhunia AK, Sain A, Chowdhury PN, et al. Text is Text, No Matter What: Unifying Text Recognition using Knowledge Distillation. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Published online October 2021: 963-972. doi: 10.1109/iccv48922.2021.00102.
30. Jaderberg M, Simonyan K, Vedaldi A, Zisserman A. Synthetic data and artificial neural networks for natural scene text recognition. arXiv. 2014, arXiv:1406.2227. 2014: 1-10. doi: 10.48550/arXiv.1406.2227
31. Karatzas D, Gomez-Bigorda L, Nicolaou A, et al. ICDAR 2015 competition on Robust Reading. 2015 13th International Conference on Document Analysis and Recognition (ICDAR). Published online August 2015: 1156-1160. doi: 10.1109/icdar.2015.7333942
32. Borisyuk F, Gordo A, Sivakumar V. Rosetta: Large scale system for text detection and recognition in images. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Published online July 19, 2018: 1-9. doi: 10.1145/3219819.3219861
33. Qiao Z, Zhou Y, Yang D, et al. SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition. in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Published online June 2020: 1-10. doi: 10.1109/cvpr42600.2020.01354