2D Positional Embedding-based Transformer for Scene Text Recognition

Zobeir Raisi — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Mohamed A. Naiel — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Paul Fieguth — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Steven Wardell — ATS Automation Tooling Systems Inc., Cambridge, ON, N3H 4R7, Canada
John Zelek — Vision and Image Processing Lab, University of Waterloo, ON, N2L 3G1, Canada
Email: {zraisi, mohamed.naiel, pfieguth, jzelek}@uwaterloo.ca, swardell@atsautomation.com

Abstract

Recent state-of-the-art scene text recognition methods are primarily based on Recurrent Neural Networks (RNNs), however, these methods require one-dimensional (1D) features and are not designed for recognizing irregular-text instances due to the loss of spatial information present in the original two-dimensional (2D) images. In this paper, we leverage a Transformer-based architecture for recognizing both regular and irregular text-in-the-wild images. The proposed method takes advantage of using a 2D positional encoder with the Transformer architecture to better preserve the spatial information of 2D image features than previous methods. The experiments on popular benchmarks, including the challenging COCO-Text dataset, demonstrate that the proposed scene text recognition method outperformed the state-of-the-art in most cases, especially on irregular-text recognition.

1 Introduction

Scene text recognition aims to convert detected text or an image patch of words into characters or words. Since the properties of scene text will generally be quite different from those of scanned documents, it is difficult to develop an effective text recognition method based on classical OCR or handwriting recognition methods, such as [1–3]. This difficulty stems from images captured in the wild including various challenging conditions [4] such as images of low resolution [5, 6], lightning extremes [5, 6], environmental conditions [7, 8], fonts [7–9], orientation [9], languages [10] and lexicons [5, 6].

Scene text recognition methods [11–16] have mainly utilized deep convolutional neural networks (DCNNs) [17, 18] and Recurrent Neural Networks (RNNs) [19], frameworks that are inspired from natural language processing. Some methods [12–14] have used connectionist temporal classification (CTC) [20], and others [15, 16], adopted an attention mechanism [21] for the prediction of character sequences. Although these methods [12–14] achieved good performance on regular-text datasets, containing primarily examples of horizontal text, their accuracy declines on irregular text datasets [6, 7, 9, 22] that contain curved and multi-oriented text.

Several recent methods [11, 12, 15, 16] have attempted to overcome the irregular-text challenge using rectification [23, 24]. For instance, Shi *et al.* [11, 15] proposed a text recognition system that combined attention-based sequence and a STN module to rectify irregular text, followed by a RNN for word recognition. However, the resulting recognition accuracy remains far from expectations.

The Transformer's architecture [25] is a novel framework, introduced first for natural language processing (NLP), taking advantage of both convolutional neural networks (CNNs) and RNNs. The architecture is less sensitive to the position of input sequences, compared to RNN and LSTM frameworks that contain inductive bias [26], because position information is not inherently encoded among the input set of sequences. The specific reason is that the self-attention and feed-forward network (FFN) layers used in Transformer make it permutative equivalent, i.e., it computes the output of each element in the input sequence independently. Although the 1D Positional Encoding (PE) technique used in Transformer [25] is able to address the permutation equivalent problem that may exist in natural language processing related 1D sequences, it is not capable of capturing the horizontal and vertical features generated by the CNNs for a 2D input image [27].

In this paper, we first extend the transformer architecture of [25] to be applicable to the recognition of 2D text images without relying on text rectification. To that end, in our proposed method we adopt a generalization of the original transformer's 1D encoding [25] to be applicable for the 2D image feature by extending the positional encoder from 1D to 2D. Experimental results show that the proposed scene text recognition architecture provides higher accuracy than that of the state-of-the-art techniques on seven out of eight challenging datasets.

2 Background

Transformer's architecture has been initially introduced in [28] for machine translation by using a new attention-based mechanism. This architecture introduces self-attention layers, which scan through each element of a sequence and update it by measuring the relationship between this element and the whole sequence [28]. The main advantages of attention-based models in transformer are their parallel computations suitability at lower memory cost, which makes them more suitable than recurrent neural networks (RNNs) [19, 29] on learning from long sequences. This transformer architecture [28] has been later exploited in natural language processing (NLP) [30, 31] and it has been recently integrated in several successful applications in speech recognition [32] and computer vision [33–35].

By dropping the PE layer, the Transformer's architecture can be viewed as a stack of N blocks ($B_n | n = 1, 2, ..N$), which each block consists of a self-attention $A_n(\cdot)$ and Feed-Forward $F_n(\cdot)$ layers. The *self-attention* layer, the key defining part of Transformer, is a normal attention block that allows the model to learn and access information of the past hidden layers. Let $x = [x_1, x_2, ..., x_t]^\top \in \mathbb{R}^{t \times d}$, within t and d denote the length and dimension of the input sequence. Each row of the self-attention function $A_1(x)$ can be demonstrated as a weighted sum of the value matrix $V \in \mathbb{R}^{t \times d}$, with the weights determined by similarity scores between the key matrix $K \in \mathbb{R}^{t \times d}$ and query matrix $Q \in \mathbb{R}^{t \times d}$ as follows:

$$
\begin{aligned}
&A_1(x) = \texttt{Softmax}\Big(\frac{QK^\top}{\sqrt{d}}\Big)V, \\
&Q = [q_1, q_2, ..., q_t]^\top, \qquad q_i = W_q x_i + b_q, \\
&K = [k_1, k_2, ..., k_t]^\top, \qquad k_i = W_k x_i + b_k, \\
&V = [v_1, v_2, ..., v_t]^\top, \qquad v_i = W_v x_i + b_v,
\end{aligned} \tag{1}
$$

where W and b are the weight and bias parameters introduced in $A_1(\cdot)$. As seen in Figure 1, Rather than only computing the attention once, the *multi-head mechanism* runs through the scaled dot-product attention in (1) multiple times in parallel.

3 Methodology

Figure 1 illustrates the proposed architecture for scene text recognition that inherited from the standard Transformer's architecture [25]. We can categorize it into two main modules: encoder and decoder. The main role of the encoder is to extract high-level 2D feature representations of an input image, and the decoder is used to convert these feature maps to a sequence of characters.

3.1 Encoder

The proposed encoder module utilizes the multi-head self-attention mechanism presented in Section 2, as well as three main sub-blocks that are as follows: (a) CNN Feature Extraction, (b) Spatial 2D-Positional Encoding, and (c) Feed-forward network (FFN), which can be described as follows.

CNN Feature Extraction: A CNN first processes the input image to extract a compact feature representation and learn a 2D representation of an input image. We adopt a modified ResNet-31 architecture [18] for the CNN backbone. During implementation, all the input images are converted into grayscale and resized to 32×100 pixels.

Spatial 2D-Positional Encoding: Transformer is permutation equivalent [36], so some extra care is required to retain the 2D structure of the image. To that effect, in our proposed models we

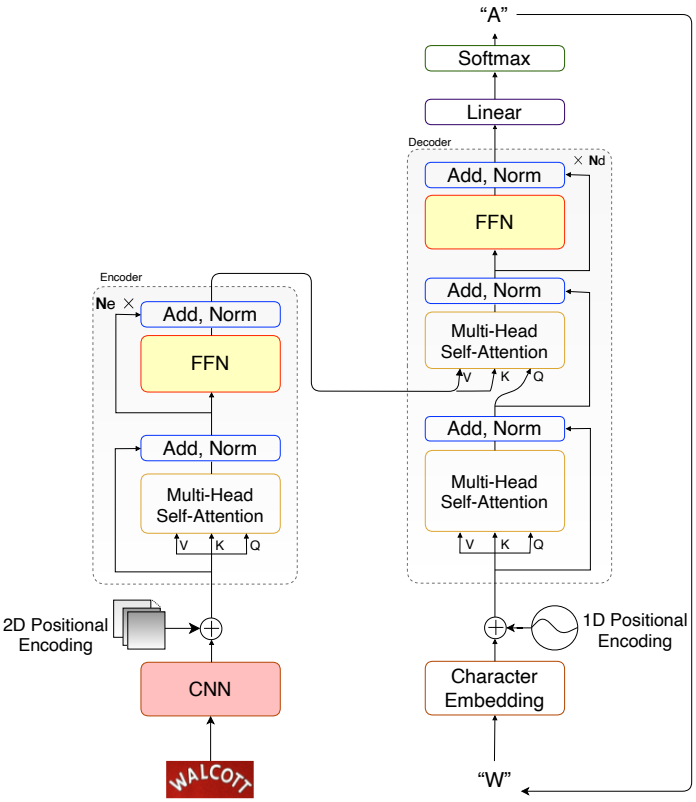

Fig. 1: The proposed text recognition using transformer architecture, where N_e and N_d denote the number of layers in the encoder and decoder. Unlike [25], our proposed architecture utilizes 2D positional encoding, a ResNet-31 backbone and a FFN layer.

adopt a generalization of the original transformer's 1D encoding [25] to be applicable for the 2D image feature by adding a fixed 2D positional encoding $\Phi(\cdot)$ made of sinuses of different frequencies as follows:

$$\begin{aligned}
\Phi(x,y,2i) &= \sin\left(x \cdot c^{4i/d}\right), \\
\Phi(x,y,2i+1) &= \cos\left(x \cdot c^{4i/d}\right), \\
\Phi(x,y,2j+d/2) &= \sin\left(y \cdot c^{4j/d}\right), \\
\Phi(x,y,2j+1+d/2) &= \cos\left(y \cdot c^{4j/d}\right).
\end{aligned} \tag{2}$$

where $c = 10^{-4}$, x and y specify the horizontal and vertical positions, and $i, j \in [0, d/4]$ and d denotes the dimension. The PE signals in (2) are added to the 2D feature outputs of the CNN block in Figure 1. We then concatenate the encoded CNN features to get the final d-channel positional encoding.

Feed-Forward Network (FFN) Layers: Here, we used a modified version of the FFN layer in the original transformer [25] to make it more robust in capturing the features generated by the encoder's multi-head self-attention mechanism. The modified FFN consists of 2 layers of 1×1 convolutions with ReLU activations [37] followed by a residual connection after the 2 layers.

3.2 Decoder

The decoder module we used follows the standard architecture of the transformer in [25]. Its main role is to use an autoregressive model to predict the decoded sequence of characters by attending the visual features generated by the encoder to predict the next sequence of characters.

4 Experimental Results

In this section, we present an experimental evaluation for the proposed method and a selected state-of-the-art scene text recognition [11–16] techniques on recent public datasets [5–9, 22, 38] that include wide variety of challenges.

Datasets: We use two type of datasets for evaluating the recognition results (1) regular-text recognition datasets: IIIT5k [5], SVT [39], ICDAR03 [38] and ICDAR13 [8] that mainly contain horizontal text, and (2) irregular-text recognition datasets: ICDAR15 [7], SVT-P [22], CUT80 [9] and COCO-Text [6], which contain multi-oriented and curved text, and these datasets are more challenging than regular datasets.

Evaluation Metrics: Word recognition accuracy (WRA) is a commonly used evaluation metric, due to its application in our daily life instead of character recognition accuracy, for assessing the text recognition schemes [11–13, 15, 16]. Given a set of cropped word images, WRA is defined as follow:

$$\text{WRA (\%)} = \frac{\text{No. of Correctly Recognized Words}}{\text{Total Number of Words}} \times 100 \tag{3}$$

Quantitative Results: By applying the proposed architecture in Figure 1 on benchmark datasets, we trained our model on 36 classes of characters. Table 1 shows a comparison in terms of the WRA for the proposed method vs the methods in consideration. As seen from this table, the proposed model achieves competitive WRA results compared to most of the state-of-the-art methods in different datasets. For three regular-text dataset, SVT [39], ICDAR03 [38] and ICDAR13 [8], it outperformed all the state-of-the-art methods with WRA of 89.34%, 95.85% and 93.89%, respectively. Furthermore, it achieved the best performance on SVT-P [22] dataset, which contains only curved and irregular text, with a large margin of 4% compared to all other methods. This performance demonstrates the strength of the Transformer network in recognizing arbitrary shapes of text compared to RNN-based methods even without using a text rectification module. Unlike the recent Transformer-based scene text recognition method in [27] that depends on ResNet-101 and adaptive 2D PE module, the proposed method utilizes a lighter backbone architecture, namely ResNet-31, and fixed 2D PE as can be seen in (2).

Qualitative Results: Figure 2 shows the qualitative performances for the proposed methods on some failure cases that provide by [16], which all the methods in [11–16] failed on these images. As shown in Figure 2(a), the proposed method recognized correctly all these images that mostly contain irregular text. We also show some failure cases of our proposed model in Figure 2(b), which it mainly failed in images that contain occluded characters. It worth noting that despite the proposed scheme offers "guice" for the sample labelled as "guide", this partially occluded example looks for several humans as "guice" as well, which indicates a noisy ground truth data and the robustness of our model to partial-occlusion.

5 Conclusion

In this paper, we have presented a new scene text recognition architecture based on integrating a 2D positional encoder with the Transformer. Furthermore, we have proposed a new feed-forward-network layer in the encoder module to make it more robust in capturing the features generated by the encoder's self-attention mechanism. The new proposed scene text recognition architecture better preserves the spatial information in 2D image features than the prior methods. Experimental results on eight public datasets have demonstrated that the proposed scene text recognition method has offered higher WRA than six recent state-of-the-art RNN-based models in most of the cases, specially on irregular-text recognition datasets. Since the Transformer's architecture requires more computations than RNN-based frameworks in the inference time, we would like to optimize the speed of the proposed Transformer's to make it faster in scene text recognition.

Acknowledgments

The authors would like to thank the Ontario Centres of Excellence (OCE) - Voucher for Innovation and Productivity II (VIP II) - Canada program, and ATS Automation Tooling Systems Inc., Cambridge, ON Canada, for supporting this research work

References

[1] H. Bunke and P. S.-p. Wang, *Handbook of character recognition and document image Anal.* World scientific, 1997.

Table 1: Comparing the WRA of some of the recent text recognition techniques using IIIT5k [5], SVT [39] , ICDAR03 [38], ICDAR13 [8], ICDAR15 [7], SVT-P [22], CUT80 [9] and COCO-Text [6] datasets. Best and second best methods are highlighted by bold and underline, respectively.

Method	IIIT5k	SVT	ICDAR03	ICDAR13	ICDAR15	SVT-P	CUT80	COCO-Text
CRNN [13]	82.73%	82.38%	93.08%	89.26%	65.87%	70.85%	62.72%	48.92%
RARE [11]	83.83%	82.84%	92.38%	88.28%	68.63%	71.16%	66.89%	54.01%
ROSETTA [14]	83.96%	83.62%	92.04%	89.16%	67.64%	74.26%	67.25%	49.61%
STAR-Net [12]	86.20%	86.09%	94.35%	90.64%	72.48%	76.59%	71.78%	55.39%
CLOVA [16]	87.40%	87.01%	94.69%	92.02%	75.23%	80.00%	74.21%	57.32%
ASTER [15]	**93.20%**	89.20%	92.20%	90.90%	74.40%	80.90%	81.90%	60.70%
Proposed	89.23%	**89.34%**	**95.85%**	**93.89%**	**75.78%**	**84.34%**	**84.03%**	**65.80%**

Fig. 2: Qualitative results of the proposed method, which in (a) shows the correctly predicted, and (b) illustrates the failure cases. It must be noted that all the methods in [11–16] have been failed on the above images.

[2] J. Zhou and D. Lopresti, "Extracting text from www images," in *ICDAR*, vol. 1, 1997, pp. 248–252.

[3] N. Arica and F. T. Yarman-Vural, "An overview of character recognition focused on off-line handwriting," *IEEE Trans. on Syst. , Man, and Cybernetics, Part C (Appl. and Reviews)*, vol. 31, no. 2, pp. 216–233, 2001.

[4] Z. Raisi, M. A. Naiel, P. Fieguth, S. Wardell, and J. Zelek, "Text detection and recognition in the wild: A review," *arXiv preprint arXiv:2006.04305*, 2020.

[5] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *BMVC*, 2012.

[6] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[7] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *ICDAR*, 2015, pp. 1156–1160.

[8] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras, "Icdar 2013 robust reading competition," in *ICDAR*, 2013, pp. 1484–1493.

[9] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. with Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014.

[10] M. Iwamura, N. Morimoto, K. Tainaka, D. Bazazian, L. Gomez, and D. Karatzas, "Icdar2017 robust reading challenge on omnidirectional video," in *Proc. IAPR ICDAR*, vol. 1, 2017, pp. 1448–1453.

[11] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *IEEE CVPR*, 2016, pp. 4168–4176.

[12] 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*. BMVA Press, September 2016, pp. 43.1–43.13.

[13] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *TPAMI*, vol. 39, no. 11, pp. 2298–2304, 2016.

[14] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, 2018, pp. 71–79.

[15] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "Aster: An attentional scene text recognizer with flexible rectification," *TPAMI*, 2018.

[16] 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," in *ICCV*, 2019.

[17] B. Su and S. Lu, "Accurate scene text recognition based on recurrent neural network," in *ACCV*. Springer, 2014, pp. 35–48.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE CVPR*, pp. 770–778, 2015.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[20] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *ICML*, 2006, pp. 369–376.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[22] T. Quy Phan, P. Shivakumara, S. Tian, and C. Lim Tan, "Recognizing text with perspective distortion in natural scenes," in *ICCV*, 2013, pp. 569–576.

[23] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," in *Proc. Int. Conf. on Neural Inf. Process. Syst. - Volume 2*. MIT Press, 2015, pp. 2017–2025.

[24] F. Zhan and S. Lu, "Esir: End-to-end scene text recognition via iterative image rectification," in *IEEE CVPR*, 2019, pp. 2059–2068.

[25] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[26] X. Liu, H.-F. Yu, I. Dhillon, and C.-J. Hsieh, "Learning to encode position for transformer with continuous dynamical model," *arXiv preprint arXiv:2003.09229*, 2020.

[27] J. Lee, S. Park, J. Baek, S. Joon Oh, S. Kim, and H. Lee, "On recognizing texts of arbitrary shapes with 2D self-attention," in *IEEE CVPR*, 2020, pp. 546–547.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008. [Online]. Available: https://arxiv.org/pdf/2005.12872.pdf

[29] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[32] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[33] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, and D. Tran, "Image transformer," *arXiv preprint arXiv:1802.05751*, 2018.

[34] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 244–253.

[35] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," *arXiv preprint arXiv:1805.08318*, 2018.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," *arXiv preprint arXiv:2005.12872*, 2020.

[37] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.

[38] S. M. Lucas, A. Panaretos, L. Sosa, A. Tang, S. Wong, and R. Young, "Icdar 2003 robust reading competitions," in *ICDAR, 2003. Proceedings.*, Aug 2003, pp. 682–687.

[39] K. Wang and S. Belongie, "Word spotting in the wild," in *Proc. Eur. Conf. on Comp. Vision*. Springer, 2010, pp. 591–604.