End-to-End Transformer-Based Architecture for Text Recognition from Document Images

Dipankar Ganguly, Akkshita Trivedi, Bhupendra Kumar, Tushar Patnaik, and Santanu Chaudhury

1 Introduction

Towards development of robust Optical Character Recognition System (OCR) with high tolerance to image degradation as well as capacity to handle deprecated characters, several challenges are faced. One of the reasons attributed to this is non-availability of such noisy datasets to fine tune systems. Furthermore, the standard evaluation methods are comprised of high quality datasets. Thus, end-to-end realization of OCRs with such capacity is seldom encountered even within the research community.

Motivation of our work derives from the end goal of having an End-to-End book conversion pipeline, agile enough to suit the peculiarities of several books still possess strong recognition accuracies. Recurrent Neural Networks (RNNs) [7] or specifically the Bi-Directional Long Short-Term Memory Sequences (Bi-LSTMs) [18] in document images have fairly established the state of the art in terms of End-to-End Modelling. Nevertheless, the models are powerful and comprehend the learning of diverse scripts and language peculiarities. Considerable efforts have been made in sequence-to-sequence modelling, particularly in language models and Machine Translation [4, 6, 38], thus re-shaping the encoder-decoder architectures [26, 41]. More recently, the Transformer-based approaches have gained momentum [39], though the very limited application of the same is seen in text recognition.

Supported by MeitY-Ministry of Electronics & Information Technology, Government of India.

D. Ganguly (✉) · B. Kumar · T. Patnaik
Centre for Development of Advanced Computing (C-DAC), Noida, Uttar Pradesh, India
e-mail: dipankargangulycdac@gmail.com

A. Trivedi · S. Chaudhury
Indian Institute of Technology (IIT), Jodhpur, Rajasthan, India

U. Mudenagudi et al. (eds.), *Proceedings of the Satellite Workshops of ICVGIP 2021*,
Lecture Notes in Electrical Engineering 924,
https://doi.org/10.1007/978-981-19-4136-8_10

The standard Recurrent networks, particularly LSTMs perform computation or tuning alongside the symbol positions from the input and output. This kind of sequential nature inhibits any parallelization of training, this has also been discussed by others [26, 39]. Thus, training longer sequences impose extremely long training times. In terms of computational efficiency, various approaches have been proposed and significant work has been carried out [1]. However, inherent sequential nature of the problem still remains with all the approaches.

On the other hand, Attention-based mechanisms have consistently secured their place in sequence-to-sequence modelling tasks [12]. The Attention mechanism essentially enables the global dependencies between the input and output. However, Attention-based mechanisms are mostly combined with RNN architectures. Thus, the shortcoming of the existing architecture motivated design and experiment with our existing architecture particularly for historical document recognition.

In this paper, we propose a Transformer-based OCR architecture fused with Masked BERT Language Model with attention layers for document image recognition in addition to GAN and DBPN-based denoising and Super-Resolution (SR) [32] for state-of-the-art document image recognition. We shall share our experiences in designing the architecture particularly in relation to convergence of Transformer Network and the insights of our experiment. We empirically show that the architecture outperforms existing state of the art for Odiya document images from Odia Virtual Academy [29].

2 Background

The standard process of text recognition for document images begins with denoising and Super-Resolution (SR) of document image [8, 14, 44]. Denoising becomes an essential component as high noise is observed in historical documents. Subsequently, SR the module intends to derive a high resolution (HR) output from a low resolution (LR) document image. The challenges like low resolution and the presence of skew are being handled within this layer.

This is followed by Image Document segmentation which is intended to extract text segments comprising either a character or word or even a line. Essentially, text-segmentation works by labelling a set of spatially adjusted features comprising of group of pixels with visual highlights. Broadly, segmentation is classified as line-based, word based or character-based [2, 35].

The latest developments for recognition are centred around end-to-end memory networks, which are usually based on a recurrent attention mechanism, which are proven to be better in comparison with sequence-aligned recurrence across various tasks. This module finally takes the segmented sequences and generates corresponding characters or words. In our proposed architecture, the segmented words are fed to our Transformer-based model architecture fused with a Masked BERT Language Model to recognize text. This has been coupled with Global and Normalized Attention Mechanisms, enabling transformers to support more parallelism with faster convergence and state-of-the-art accuracies.

3 Related Work

There are a handful of works on Transformer architecture-based Text Detection or Object Detection, such by Lyu et al. [27]. They proposed an architecture for Scene Text Detection using Attention and Transformer-based units to recognize texts. In another work [5] Carion et al. have used Transformers in Object Detection and have presented a novel framework.

In direction of sophisticated SR, due to breakthrough advancement in the field of the deep neural networks, many new SR methods have been proposed [11, 20, 21]. Similarly, Zhang et al. [43] presented a CNN-based Super resolution algorithm specific to the demands of OCR. However, the proposed methods are mostly feed-forward in nature. Thereby having inability of mapping the relationship between LR and HR. Hence, in this paper, we try to integrate SR method proposed by Haris et al. [15], which proposes a projection network, feeding back the error predictions at each layer, collecting the self-correcting features for up-sampling at every stage to improve SR resolution.

In terms of Segmentation, Long et al. [24] revolutionized the idea of semantic segmentation by using ImageNet database [9]. They have up-sampled the images with deconvolutional layers and subsequently appending lower layers to improve predictions. In other domains such as Scene text detection, in design of image segmentation, many deep learning-based algorithms have been developed with promising results [23, 25, 42]. In document image segmentation, Ronneberger et al. [33] introduced U-Net by proposing a U-shaped symmetric architecture between the contracting path to capture setting and expanding path that empowers exact restrictions, with a lower layer for each level. Also, Santos et al. [34] developed a multi-step handwritten text-segmentation framework, utilizing Y and X histogram projections to eliminate false lines and words, respectively.

Hochreiter and Schmidhuber [19] introduced Long Short-Term Memory in 1997, which has become de-facto standard in the Text Recognition. Graves et al. [13] proposed the use of Bi-Directional LSTM(Bi-LSTM) architecture which allowed bi-directional (forward and backward layers) longer range context. In this paper, we propose integrating the architecture introduced by Ray et al. [31] which combines Connectionist Temporal Classification(CTC) to learn the labelling unsegmented and unaligned sequence of data and a Bi-LSTM model for the advantages mentioned above.

Developments of complete text recognition solutions, such as [28] proposed Arabic handwritten document segmentation framework, utilizing a variant of U-net with residual blocks(RU-net) for text line segmentation; for word segmentation, BLSTM-CTC (Bidirectional Long Short-Term Memory followed by a Connectionist Temporal Classification). In the similar line, CRAFT [3] proposed a text detector without character annotations by generating a pseudo character-level ground truths from an interim word-level datasets. Also, EAST [45] was designed for fast and accurate text detection with a single neural network and with appropriate loss function, rotated

rectangular or quadrilateral text regions are generated. In another work by Liao et al. [22], a single layer neural network word-based text detector called TextBoxes, combined with a text recognition algorithm named Convolutional Recurrent Neural Network(CRNN) [36] produced significant results in terms of prediction.

Significant accuracy in text-based detection is also discussed in [40] by introducing a deep learning-based super-resolution framework without additional computing cost.

4 Proposed Framework

Our proposed architecture comprises cascaded document image enhancement and recognition as depicted in Fig. 1. It broadly comprises of three modules, i.e. denoising and super-resolution module, U-net-Based Segmentation, and Transformer-based Recognition Combined with BERT Language Model as decoder which is jointly optimized.

Initially, the document image is fed into the DBPN [15] and GAN module which achieves denoising and SR task by an iterative, multi-stage up (extracting features to upgrade to HR) and down (resizing the image as per LR configuration)-sampling operators, connected mutually to extract the non-linear relation between LR and HR, as shown in the Fig. 1.

The U-net [33] architecture has been modified, which consists of 2 paths, the encoder or the contracting path catches the settings in the image. The encoder consists of convolutional and max-pooling layers stacked together. The decoder or the expanding path empowers exact limitation utilizing transposed convolutions. Thus, an end-to-end FCN is generated, which can accept an input image of any size. Figure 1 shows the U-Net architecture.

Finally coming to recognition, the proposed transformer architecture consists of a Residual Network [16] (Res-Net) layer at first, particularly the pre-trained Res-Net 18 is used with Transfer Learning approach [30]. The transfer learned Res-Net 18 is followed by a fully connected layer that acts as an bridge to connect the encoder transformer units, which is a modified Multi-Head Attention [39] and is a multi-layer bi-directional transformer encoder. The layer is responsible to capture global dependencies between the feature map and output, this works by aggregating information from the parts of the input. This layer is followed by a Normalized Attention Layer that acts as a separator to disconnect output with attention computation, thus enabling parallel and faster optimization and convergence. Finally a word-based decoder is fused with BERT Masked LM to decode corresponding words.

Our approach is mostly script independent and would work for a wide variety of historical document images. We have performed our experiments on Odia Historical books from Odia Virtual Academy as such datasets are not popular and very difficult to curate.

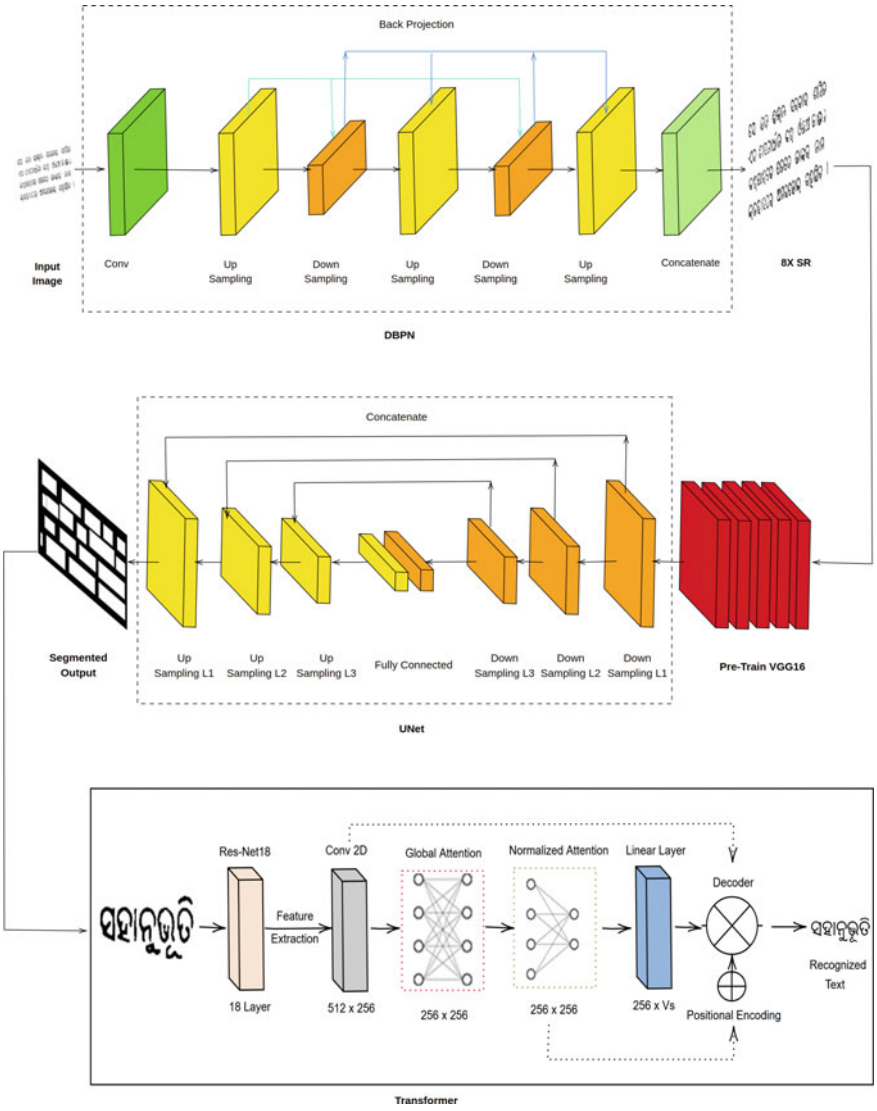

Fig. 1 The proposed Transformer Architecture with Masked BERT LM decoder

5 Methodology

In this section, we describe finer details about our proposed framework.

5.1 Super Resolution and Segmentation

Document image size of 64×64 is fed to the DBPN architecture at first, the image size is up-scaled to 512×512 and then, back-projected to 64×64, the residual between the reconstructed and the observed LR/HR maps is up/down-scaled accordingly and added with the observed maps to enhance the extraction. The result is a SR image of output 512×512. The SR image picks up pre-trained weights from VGG16 model, such that we eliminate the possibility of weights being assigned randomly, thus increasing our chances of accurate prediction. The pre-trained weights and SR image is now passed to U-Net. During encoding, the SR image undergoes 2 un-padded convolutions recurrently, with each convolution followed by a leaky Rectified Linear Unit (ReLU) and max-pooling for down-sampling with a stride of 2. The number of feature channels doubles at every step of down-sampling ($64 \rightarrow 128 \rightarrow 256 \rightarrow 512$). During decoding, the feature maps are up-sampled, followed by halving the feature channels and adding with the corresponding down-sampled features at the time of encoding, and 2 un-padded convolutions and leaky ReLU. The output from U-Net, i.e. the extracted words are then fed to the proposed transformer. The segmented words are then fed to the transformer.

5.2 Global Attention

In order to capture global dependencies, we have used a Global Attention mechanism which is built with Multi-Head Attention units [39] which works to aggregate information from input. Instead of going linearly performing single attention, we perform this in parallel by projecting the queries, keys, and values of the transformer to different dimensions across each value. We apply global attention module to the flattened feature map generated by Res-Net followed by a convolutional layer, the feature sequence I with shape of $k \times c$, which precisely in our case is 512×256. Corresponding to each feature vector we use $I_i (i \in [1, k])$,the position vector E_i is embedded with position index i. Subsequently, a feature vector F with position embedded information is available. We have incorporated many transformer layers into series to aggregate information from F. Within each transformer unit, the corresponding query, keys and values are obtained as [27]:

$$Q_l^i = \begin{cases} F_i & l = 1, \\ O_{l-1}^i & l > 1 \end{cases} \tag{1}$$

$$K_l^i = \begin{cases} F_i & l = 1, \\ O_{l-1}^i & l > 1 \end{cases} \tag{2}$$

$$V_l^i = \begin{cases} F_i & l = 1, \\ O_{l-1}^i & l > 1 \end{cases} \tag{3}$$

In the Eq. 1, Q_l^i represents query vector for the $i-$th transformer unit in the l_th transformer layer. In same equation O_{l-1}^i represents the output from the previous transformer layer. Alongside Q_l^i, in Eq. 2 K_l^i & 3 V_l^i are corresponding key and value vectors for the same l_th layer of the transformer.

With all necessary query, key and value vectors computed within each transformer unit, the overall transformer output is weighted summation of all the values. The weights sum of each layer is calculated with the formula :

$$\alpha_l^{ij} = \frac{\exp(W_l^q Q_l^i . W_l^k Q_l^{ij})}{\sum_{j'=1}^{k} \exp(W_l^q Q_l^i . W_l^k Q_l^{ij})} \tag{4}$$

In the Eq. 4, W_l^q denotes the trainable weights across the layers. Finally, the output of each transformer is represented as

$$O_l^i = \text{act\_func}(\sum_{j=1}^{k} \alpha_(W_l^v . V_l^{ij}) \tag{5}$$

In the above Eq. 5, W_l^v is the learned weight and act_func is a non-linear activation which is can be referred from [10, 39]. The outputs of the last transformer is taken as the global attention layer output.

5.3 Normalized Attention

Normalized Attention is a sparse layers of transformer encoders to aid parallelism in attention and provide separation of Global Attention to Outputs. Residual connections are built to promote information exchange between the layers.

The basic attention, as described in [37], is designed to work serially and are usually integrated with Recurrent Networks as

$$\alpha_t = \text{Attention}(h_{t-1}, \alpha_{t-1}, \text{I}) \tag{6}$$

In the above Eq. 6, h_{t-1} and α_{t-1} represents the weights of the hidden state and attention of the previous steps of RNN decoder, I is the encoded feature sequence. Therefore, computation of the step t is limited by previous steps, which is a bottleneck.

To overcome this problem, we have devised a Normalized Attention, by modifying the architecture proposed in [27]. In this layer, the dependency relationships are not fully dependent and can be optimized simultaneously. The normalization is dependent on the following function:

$$\alpha_t = \text{softmax}(W_2 \tanh(W_1 O^T)) \tag{7}$$

In the above equation W_1 W_2 represents learnable parameters on the global attention layer.

Once the computation is completed only the normalized weights coefficients α, encoded within the feature sequence are obtained as output from the layer, which is given by the Eq. 8, where indexes i & j represent the outputs node and feature vector index:

$$G_i = \sum_{j=1}^{k} \alpha_{ij} I_j \tag{8}$$

This layer prevents the range of values in the other layers into changing too much, thus the model trains faster and has a better ability to generalize.

All the available OCRs are compared with our proposed architecture and compared to empirically show the best with Critical distance diagram. In Fig. 2 character recognition accuracies are compared whereas Fig. 3 showcases comparison with word recognition accuracies. The abbreviations are OCR EG—Classical Odia OCR based on cascaded rule-based architectures; DNN—Deep Bi-LSTM with CTC Recognition Engine; Tesseract OCR for Odia; DNN LM—Deep Bi-LSTM with CTC Recognition with Fraternal dropout-based Language Model; TOCR—Proposed Transformer-based OCR

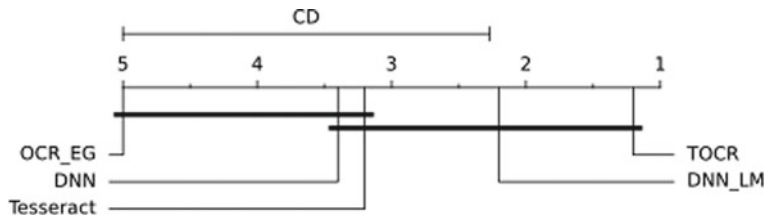

Fig. 2 Character Level Critical Distance

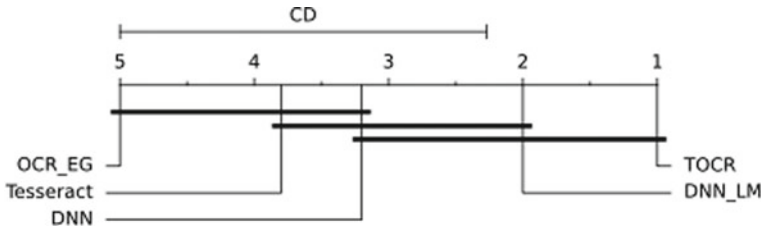

Fig. 3 Word Level Critical Distance

5.4 Decoder

The transformer decoder is fused with a pre-trained Language model on Odia Corpus. We have 1 Million Odia sentences from Odia Virtual Academy Books and Web Crawl. Thereby, we have trained a Masked BERT LM [10]. The decoder of the transformer unit is combined with the Masked LM decoder and predicts output words with probability given by

$$P_i = \text{softmax}(WG_i + b) \tag{9}$$

Here, W and b are weights and biases learned during training.

This amalgamation of Masked LM is essential as the Normalized Attention removes certain dependencies which are pivotal to the decoder and is compensated by the BERT LM. The entire transformer network is jointly optimized (Table 1).

6 Experimental Results

6.1 Dataset

We have tested our approach on Odia script. We have 26 books from Odia Virtual Academy scanned at 300 dpi, of which 17 books are very old; published between 1920 and 1980, and 9 books are relatively new (published after 1980). In total our dataset comprises 1700 pages with total 423 K words and 54 K unique words. Thus, the dataset has a good blend of very old, degraded, noisy documents along with newer printed documents. These together combine to put forth a plethora of document recognition challenges. Text graphics separation as well as segmentation is performed before the document is fed into the recognition pipeline.

Our datasets are divided into Training (80%), Test (10%), and Validation (10%). The DBPN along with SR with DBPN and U-net is trained on 15K segmented noise resilient images along with corresponding original noise present in the image.

Table 1 The Table compares the accuracy of our proposed framework with existing Odia OCR

Validation data	Proposed OCR		DNN OCR + LM		DNN OCR		OCR engine(Consortia)		Tesseract	
	Word (%)	Char (%)	Word (%)	Char(%)	Word (%)	Char (%)	Word (%)	Char (%)	Word (%)	Char (%)
Test Set 1	**92.12**	**94.03**	90.60	93.80	89.00	93.50	68.84	86.78	85.70	93.52
Pages: 89										
Words: 23457										
Test Set 2	**90.15**	**93.34**	86.83	91.23	85.00	90.91	7.64	12.30	20.03	53.47
Pages: 34										
Words: 8302										
Test Set 3	**88.05**	**94.79**	84.74	91.79	83.00	91.03	3.47	7.40	15.93	47.24
Pages: 26										
Words: 6584										
Test Set 4	**87.12**	**93.81**	83.05	92.61	82.00	92.30	5.40	16.66	17.24	52.50
Pages: 103										
Words: 23893										
Test Set 5	**89.78**	**92.05**	83.47	91.32	82.35	91.18	70.63	87.13	83.22	93.65
Pages: 94										
Words: 25690										

6.2 Results

We have performed intensive testing across different available OCRs for Odia and also carried out hypothesis testing using Autorank [17] for character and word-level accuracies as shown in Fig. 2.

References

1. Anwar S, Hwang K, Sung W (2015) Fixed point optimization of deep convolutional neural networks for object recognition. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp 1131–1135. IEEE
2. Arica N, Yarman-Vural FT (2001) An overview of character recognition focused on off-line handwriting. IEEE Trans Syst Man Cybern Part C (Appl Rev) 31(2):216–233. https://doi.org/10.1109/5326.941845
3. Baek Y, Lee B, Han D, Yun S, Lee H (2019) Character region awareness for text detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9365–9374
4. Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

5. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision, pp 213–229. Springer
6. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078
7. Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555
8. Dai D, Wang Y, Chen Y, Gool LV (2015) How useful is image super-resolution to other vision tasks? CoRR arXiv:1509.07009
9. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp 248–255. IEEE
10. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805
11. Dong C, Loy CC, Tang X (2016) Accelerating the super-resolution convolutional neural network. In: European conference on computer vision, pp 391–407. Springer
12. Dutil F, Gulcehre C, Trischler A, Bengio Y (2017) Plan, attend, generate: planning for sequence-to-sequence models. arXiv preprint arXiv:1711.10462
13. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J (2009) A novel connectionist system for unconstrained handwriting recognition. IEEE Trans Pattern Anal Mach Intell 31(5):855–868. https://doi.org/10.1109/TPAMI.2008.137
14. Greenspan H (2009) Super-resolution in medical imaging. Comput. J. 52(1):43–63. https://doi.org/10.1093/comjnl/bxm075
15. Haris M, Shakhnarovich G, Ukita N (2018) Deep back-projection networks for super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1664–1673
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
17. Herbold S (2020) Autorank: a python package for automated ranking of classifiers. J Open Source Softw 5(48):2173 (2020). https://doi.org/10.21105/joss.02173
18. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
19. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
20. Kim J, Lee JK, Lee KM (2016) Deeply-recursive convolutional network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1637–1645
21. Lai WS, Huang JB, Ahuja N, Yang MH (2017) Deep laplacian pyramid networks for fast and accurate super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 624–632
22. Liao M, Shi B, Bai X, Wang X, Liu W (2017) Textboxes: a fast text detector with a single deep neural network. Proceedings of the AAAI conference on artificial intelligence, vol 31(1). https://ojs.aaai.org/index.php/AAAI/article/view/11196
23. Liu X, Meng G, Pan C (2019) Scene text detection and recognition with advances in deep learning: a survey. Int J Document Anal Recogn (IJDAR) 22(2):143–162
24. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
25. Long S, He X, Yao C (2020) Scene text detection and recognition: the deep learning era. Int J Comput Vis, 1–24
26. Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. arXiv preprint arXiv:1508.04025
27. Lyu P, Yang Z, Leng X, Wu X, Li R, Shen X (2019) 2d attentional irregular scene text recognizer. arXiv preprint arXiv:1906.05708

28. Neche C, Belaid A, Kacem-Echi A (2019) Arabic handwritten documents segmentation into text-lines and words using deep learning. In: 2019 international conference on document analysis and recognition workshops (ICDARW), vol 6, pp 19–24. https://doi.org/10.1109/ICDARW.2019.50110
29. Parida S, Bojar O, Dash SR (2020) Odiencorp: odia–english and odia-only corpus for machine translation. In: Smart intelligent computing and applications, pp 495–504. Springer
30. Pratt LY (1993) Discriminability-based transfer between neural networks. Advances in neural information processing systems, pp 204–204
31. Ray A, Rajeswar S, Chaudhury S (2015) Text recognition using deep blstm networks. In: 2015 Eighth international conference on advances in pattern recognition (ICAPR), pp 1–6. https://doi.org/10.1109/ICAPR.2015.7050699
32. Ray A, Sharma M, Upadhyay A, Makwana M, Chaudhury S, Trivedi A, Singh A, Saini A (2019) An end-to-end trainable framework for joint optimization of document enhancement and recognition. In: 2019 international conference on document analysis and recognition (ICDAR), pp. 59–64. IEEE
33. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention, pp 234–241. Springer
34. Santos RP dos, Clemente GS, Ren TI, Cavalcanti GD (2009) Text line segmentation based on morphology and histogram projection. In: 2009 10th international conference on document analysis and recognition, pp 651–655. IEEE
35. Santos RPD, Clemente GS, Ren TI, Cavalcanti GDC (2009) Text line segmentation based on morphology and histogram projection. In: Proceedings of the 2009 10th international conference on document analysis and recognition, pp 651–655. ICDAR '09, IEEE Computer Society, USA. https://doi.org/10.1109/ICDAR.2009.183
36. Shi B, Bai X, Yao C (2016) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans Pattern Anal Mach Intell 39(11):2298–2304
37. Shi B, Wang X, Lyu P, Yao C, Bai X (2016) Robust scene text recognition with automatic rectification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4168–4176
38. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. arXiv preprint arXiv:1409.3215
39. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention is all you need. arXiv preprint arXiv:1706.03762
40. Wang L, Li D, Zhu Y, Tian L, Shan Y (2020) Dual super-resolution learning for semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)
41. Wu Y, Schuster M, Chen Z, Le QV, Norouzi M, Macherey W, Krikun M, Cao Y, Gao Q, Macherey K et al (2016) Google's neural machine translation system: bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144
42. Xu Y, Wang Y, Zhou W, Wang Y, Yang Z, Bai X (2019) Textfield: learning a deep direction field for irregular scene text detection. IEEE Trans Image Process 28(11):5566–5579. https://doi.org/10.1109/TIP.2019.2900589
43. Zhang H, Liu D, Xiong Z (2017) Cnn-based text image super-resolution tailored for ocr. In: 2017 IEEE visual communications and image processing (VCIP), pp 1–4. https://doi.org/10.1109/VCIP.2017.8305127
44. Zhang L, Zhang H, Shen H, Li P (2010) A super-resolution reconstruction algorithm for surveillance images. Signal Process 90(3):848–859
45. Zhou X, Yao C, Wen H, Wang Y, Zhou S, He W, Liang J (2017) East: an efficient and accurate scene text detector. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5551–5560