A Comparative Study of Attention-based Encoder-Decoder Approaches to Natural Scene Text Recognition

Fuze Cong*†, Wenping Hu‡, Qiang Huo‡, Li Guo†
†Beijing University of Posts and Telecommunications
Key Lab of Universal Wireless Communications, Ministry of Education
‡Microsoft Research Asia, Beijing, China
Email: †{congfz, guoli}@bupt.edu.cn ‡{wenh, qianghuo}@microsoft.com

***Abstract*—Attention-based encoder-decoder approaches have shown promising results in scene text recognition. In the literature, models with different encoders, decoders and attention mechanisms have been proposed and compared on isolated word recognition tasks, where the models are trained on either synthetic word images or a small set of real-world images. In this paper, we investigate different components of the attention based framework and compare its performance with a CNN-DBLSTM-CTC based approach on large-scale real-world scene text sentence recognition tasks. We train character models by using more than 1.6M real-world text lines and compare their performance on test sets collected from a variety of real-world scenarios. Our results show that (1) attention on a two-dimensional feature map can yield better performance than one-dimensional one and an RNN based decoder performs better than CNN based one; (2) attention-based approaches can achieve higher recognition accuracy than CNN-DBLSTM-CTC based approaches on isolated word recognition tasks, but perform worse on sentence recognition tasks; (3) it is more effective and efficient for CNN-DBLSTM-CTC based approaches to leverage an explicit language model to boost recognition accuracy.**

***Keywords*-Scene Text Recognition; Encoder-Decoder; Attention; Connectionist Temporal Classification**

I. Introduction

Recently, an attention-based encoder-decoder framework as illustrated in Fig. 1 has been applied to Scene Text Recognition (STR) with promising results on several benchmark tasks (e.g., [1]–[4]). It encodes an input image as a one-dimensional feature sequence or a two-dimensional feature map, attends on a specific part at each time-step, and decodes an output label sequence in an auto-regressive way. In the literature, different encoders, decoders and attention mechanisms have been investigated. However, most experiments and comparisons are conducted on isolated word recognition tasks, where models are trained with either synthetic word images (e.g., [1]–[6]) or a small set of real-world images (e.g., [7], [8]). Some issues critical to sentence recognition such as

- word segmentation errors,
- accumulated prediction errors with an increased sentence length,

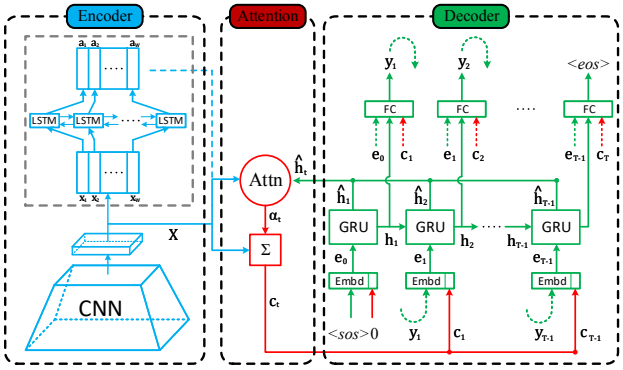


Fig. 1. Illustration of an attention-based encoder-decoder framework. The BLSTM layer plotted in the grey dashed box in *encoder* block is optional. The GRU module in *decoder* block is a dual GRU structure as in [11].

- the feasibility of leveraging an explicit language model in decoding to improve recognition accuracy,

are avoided for isolated word recognition. Consequently, observations made and conclusions drawn on isolated word recognition tasks may not be applicable to real-world sentence recognition tasks.

In this paper, we re-investigate the attention-based encoder-decoder framework and compare it with a CNN-DBLSTM-CTC based (e.g., [9], [10]) end-to-end sequence learning and labeling approach on large-scale real-world scene text sentence recognition tasks, and make the following observations:

- Compared with the CNN-DBLSTM-CTC based approach, the attention-based approach can achieve higher recognition accuracy on isolated word recognition tasks, but performs worse on sentence recognition tasks;
- It is more effective and efficient for CNN-DBLSTM-CTC based approach to leverage an explicit language model to boost recognition accuracy.

The rest of paper is organized as follows. In Section II, we describe the attention-based encoder-decoder approaches under investigation. In Section III and Section IV, we present experimental setups and results respectively. Finally, we summarize our findings and discuss future works in Section V.

* This work was done when Fuze Cong was an intern in Speech Group, Microsoft Research Asia, Beijing, China

2379-2140/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDAR.2019.00151

II. Attention-based Encoder-Decoder Approach

A general architecture of an attention-based encoder-decoder approach is shown in Fig. 1. It consists of three components: encoder, attention mechanism and decoder. Different models can be used in each of the above components.

A. Encoder

In the literature, different neural network models have been proposed to encode an input text-line image. Most of them can be grouped into two categories in terms of the shape of a learned feature map, i.e., a one-dimensional feature sequence or a two-dimensional feature map. Similar to that in machine translation, the input text-line image can be encoded as a one-dimensional feature sequence, where a spatial position in the same column shares the same attention weights, by different neural network models, e.g., a recursive Convolution Neural Network (CNN) followed by fully connection layers [1], a Convolution Recurrent Neural Network (CRNN) [2] or a self-attention network following a CNN model [12]. On the other hand, the input image can be encoded as a two-dimensional feature map to keep vertical spatial information. The attention weight corresponding to each spatial position is calculated independently. These two-dimensional attention-based approaches are widely used in irregular text recognition [4] and mathematical expression recognition [11].

In this paper, we mainly investigate whether a two-dimensional feature map is essential in regular scene text sentence recognition. Attention on a 2D feature map can provide additional vertical visual cues to improve the discrimination of ambiguous characters or symbol pairs. We first evaluate the performance of several popular CNN topologies, i.e., DenseNet99 [13], ShuffleNet50 [14] and ResNet50 [15], then compare the best one with a CRNN model by adding a bidirectional LSTM layer on top of it.

B. Attention Mechanism

Besides a conventional soft-attention mechanism [16], some advanced attention models have been proposed in scene text recognition. Cheng *et al.* [3] propose a focusing attention network to obtain more accurate alignments between an input image and an output label sequence with an extra dataset labeled on pixel level. Li *et al.* [4] improve the conventional soft-attention by taking an eight-neighborhood context into account when calculating an attention weight for each position. Zhang *et al.* [11] introduce and expand a coverage model proposed originally in [17] into an attention mechanism by considering the accumulated attention weight of each position to alleviate a miss- or over- parsing problem [18], [19]. Then, Zhang *et al.* apply this framework to handwritten mathematical expression recognition [11] and handwritten Chinese character recognition tasks [20], which achieves the state-of-the-art results.

In this paper, we compare the performance of a plain soft-attention mechanism with the one configured with a coverage model as adopted in [11] and investigate the effect of kernel size in the coverage model.

TABLE I
Overview of datasets

	Train	Val	Test	Total
G	1,019,367	54,713	12,009	1,086,089
S	323,258	44,774	36,684	404,716
Syn.	300,000	37,000	-	337,000
11Seg	-	-	19,099	19,099
IC13	-	-	1,094	1,094

C. Decoder

For the decoder module, recurrent neural network is the most widely used model. Motivated by a completely convolution-based decoder in machine translation [21], Fang *et al.* [6] propose a deep one-dimensional CNN-based decoder consisting of a dot-product attention module to capture visual cues and a character-level language module to model linguistic rules. Their experimental results show that the CNN-based decoder can achieve a comparable or slightly better accuracy than RNN-based one on several isolated word recognition tasks.

In this paper, we re-implement the CNN-based decoder proposed in [6] and compare it with a GRU-RNN based decoder on sentence recognition tasks. Similar to the decoder used in [11], a dual GRU structure is adopted here.

D. Decoding Strategy

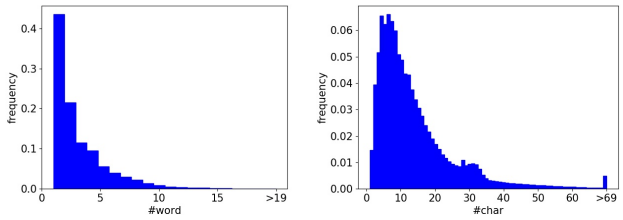
Due to the auto-regressive decoding mechanism in an attention-based system, it is difficult to parallelize decoding process across different time-steps. Beam search is widely used to obtain a decoded label sequence. To further boost recognition accuracy, some heuristic approaches have been proposed to leverage an explicit language model, e.g., N-best rescoring, shallow fusion [22], deep fusion [22] and cold fusion [23]. Considering the complexity of implementation, the N-best rescoring approach is adopted in this paper when decoding with a language model.

III. Experimental Setup

A. Datasets

Text line images used in this experiment are cropped from Open Image dataset (G for short) [24], street view images (S for short) for internal usage and synthetic text lines. The number of text lines corresponding to the training, validation and testing sets are shown in Table I. There are more than 1.6M text lines in total in the training set, which contains both isolated words and sentences. The distributions of the number of words and characters contained in each text line are shown in Fig. 2(a) and 2(b), respectively. It shows that a wide range of word length is covered, rather than single word only. Our character set consists of 26 uppercase letters, 26 lowercase letters, 10 digits, 32 punctuation and symbols, and 1 space label.

To have a solid comparison of different STR systems, a variety of testing sets are collected, including randomly



(a) # of words (b) # of characters

Fig. 2. Distributions of the number of words and characters in text lines.

sampled in-domain G and S testing sets and two out-of-domain datasets. One is an widely used academic testing set, IC13 [25] and another is collected from 11 real-world scene text recognition scenarios (11Seg for short), i.e., book cover, business card, document, Gif, invoice, menu, poster, product label, receipt, slide and street view. The numbers of lines are shown in the fourth column in Table I.

B. Training Details

The attention-based models are implemented based on PyTorch platform and trained with 32 NVIDIA Tesla P100 GPUs using a distributed synchronous gradient descent strategy [26], where the local optimizer is ADAM [27] with 0.1 times learning rate decay. The gradients of trainable parameters and the mean/variance of batch normalization (BN) are averaged over all GPUs for each iteration. Seed models with a small learning rate are trained with 10% training data on a single GPU for initialization.

Different from a conventional evaluation metric on isolated word recognition tasks [1], [3]–[6], [28], Word Error Rate (WER) and Character Error Rate (CER) are adopted on sentence recognition tasks. All the case-sensitive characters, punctuation and symbols are considered. Each word-level hypothesis is obtained by splitting a character sequence hypothesis with the assistance of the space label.

IV. Experimental Results and Discussions

A. Encoder

Firstly, we explore different backbones used in the encoder module, i.e., DenseNet99 [13], ShuffleNet50 [14] and ResNet50 [15]. The same attention mechanism and decoder topology are applied. The decoder contains 2 GRU layers, each with 256 hidden nodes and the convolution kernel size in coverage model is 11×11. The numbers of parameters and beam search (N=10) decoding results are shown in Table II. WER/CER are shown in each cell. Results show that, compared with ResNet50 and ShuffleNet50, DenseNet99 can achieve comparable or slightly better performance on different testing sets with a much smaller model size.

Secondly, we investigate whether a two-dimensional feature map is better than a one-dimensional one for attention in regular sentence recognition. Based upon DenseNet99, we add an extra BLSTM layer to further compress the two-dimensional feature maps to one-dimensional feature sequences. The BLSTM layer contains 128 hidden nodes in each direction. The results are shown in the last row in Table II. It shows that the encoder with a BLSTM layer yields worse result than the purely CNN one, though with more parameters. We conjecture that this is because the vertical spatial cues are important in recognizing some superscripts, subscripts and punctuation symbols. In following experiments, DensetNet99 is chosen as the default model in encoder.

TABLE II
WER/CER (IN %) OF DIFFERENT ENCODERS

Backbone	#Params	G	S	11Seg	IC13
ResNet50	26.7M	3.7/0.7	3.4/0.7	3.5/1.3	12.5/3.9
ShuffleNet50	7.40M	3.9/0.9	3.5/0.8	3.8/1.5	12.5/4.1
DenseNet99	5.46M	3.7/0.7	3.4/0.7	3.6/1.3	11.2/3.6
+BLSTM	11.3M	3.9/0.7	3.5/0.7	3.9/1.4	11.8/3.7

TABLE III
WER/CER (IN %) OF DIFFERENT COVERAGE MODEL

Kernel Size	#Params	G	S	11Seg	IC13
w/o	5.13M	4.0/1.0	3.8/0.9	3.8/1.6	13.8/3.7
3x3	5.40M	4.0/0.9	3.6/0.8	4.1/1.6	14.6/4.0
5x5	5.41M	3.9/0.7	3.5/0.7	3.5/1.3	11.8/3.7
7x7	5.43M	3.8/0.7	3.5/0.7	3.5/1.3	13.6/3.5
9x9	5.44M	3.8/0.7	3.5/0.8	3.7/1.4	12.6/3.7
11x11	5.46M	3.7/0.7	3.4/0.7	3.6/1.3	11.2/3.6

B. Attention Mechanism

We investigate the effect of a coverage model configured in the conventional soft-attention mechanism in scene text sentence recognition. As described in Section II, the coverage model is used to alleviate the miss- or over-parsing problem by recording which position in the feature map have been visited. The beam search results corresponding to the conventional soft-attention module and coverage models of different kernel sizes are shown in Table III. By enlarging the kernel size from 3×3 to 11×11 by stride 2, better performance can be achieved. On the other hand, compared with the plain soft-attention module, as shown in the second row in Table III, the attention model with a coverage model of 11×11 kernel size can achieve much better results. The attention mechanism with a coverage model of 11×11 kernel size is finally adopted in following experiments.

C. Decoder

Besides the GRU-RNN based decoder, we re-implement a CNN-based decoder proposed in [6], which contains 6 blocks with 1×3 convolutional kernel size and 512 channels. To make a fair comparison, the same Desnet99 based encoder is used. The results are shown in Table IV. It shows that (1) the RNN-based decoder can achieve a much lower WER consistently on different test sets than the CNN-based one; (2) The gap of CER is smaller on isolated word set (IC13) and short sentence

TABLE IV
WER/CER (IN %) OF DIFFERENT DECODERS

Decoder	#Params	G	S	11Seg	IC13
RNN	5.46M	3.7/0.7	3.4/0.7	3.6/1.3	11.2/3.6
CNN [6]	17.1M	4.1/1.0	3.7/0.8	4.0/1.8	12.3/3.7

TABLE V
WER/CER (IN %) OF ATTENTION AND CTC-BASED MODELS

	G	S	11Seg	IC13
Attention	3.7/0.7	3.4/0.7	3.6/1.3	11.2/3.6
+LM	3.4/0.6	3.0/0.6	3.3/1.2	9.7/3.3
CTC	3.4/0.7	3.4/0.7	3.2/1.2	12.7/4.0
+LM	2.9/0.5	2.7/0.6	2.7/1.1	9.4/3.1

set (S), but larger on long sentence sets (G and 11Seg). It indicates that the performance of CNN-based decoder will degrade if the length of decoding sequence is longer than a certain threshold. The RNN-based decoder is more suitable for scene text sentence recognition than the CNN-based one.

D. Comparison Between Attention and CTC-based Models

In this subsection, we compare the best attention-based encoder-decoder system (i.e., DenseNet99 based encoder, soft-attention with a coverage model of 11×11 convolution kernel size and RNN-based decoder) with one of the state-of-the-art CNN-DBLSTM-CTC based systems [29]. The CTC-based system contains a modified 10-layer VGG-Net followed by 2 BLSTM layers with 128 nodes for each direction, which is the same as that used in [29].

To boost recognition accuracy, we leverage an explicit language model in the decoding stage. In CTC-based system, a hybrid word and sub-word level bi-gram language model is used. The language model, a lexicon of 131k words/subwords and the Hidden Markov Model (HMM) topology are integrated and represented as a Weighted Finite State Transducer (WFST). The building process is similar to that described in [30]. In attention-based system, the N-best rescoring approach is adopted to leverage the language model score. The N-best hypotheses are re-ranked by the interpolated character model score and its language model score obtained from a character level, LSTM-RNN language model, which contains 2 LSTM layers with 1024 nodes for each layer.

The recognition results with or without leveraging an explicit language model are shown in Table V. Without using a language model, the attention-based system can achieve comparable performance as the CTC-based system, i.e., better on IC13 test set, the same on S test set, but slightly worse on G and 11Seg test sets. However, it performs worse than CTC-based system consistently on different test sets when leveraging an explicit language model.

To verify the effectiveness of the generated N-best hypotheses, we calculate Oracle Error Rates (OERs) on top-N beam search results, denoted as top3, top5 and top10. The OERs are obtained by selecting the hypothesis from the top-N beam search results which yields the lowest CER compared with the ground truth. Obviously, a larger N will yield a smaller oracle WER/CER. As shown in Table VI, the oracle results are much better than the best CTC and attention-based decoding results. In the future, we will investigate smarter decoding strategies or more powerful language models to better utilize the beam search N-best results to achieve better final results.

TABLE VI
ORACLE WER/CER (IN %) IN ATTENTION-BASED SYSTEM

	G	S	11Seg	IC13
Beam search	3.7/0.7	3.4/0.7	3.6/1.3	11.2/3.6
Oracle in top3	1.7/0.4	1.4/0.4	2.5/1.1	7.3/2.5
Oracle in top5	1.4/0.3	1.0/0.2	2.2/1.0	5.9/2.1
Oracle in top10	1.1/0.2	0.7/0.1	1.9/0.9	5.3/1.8

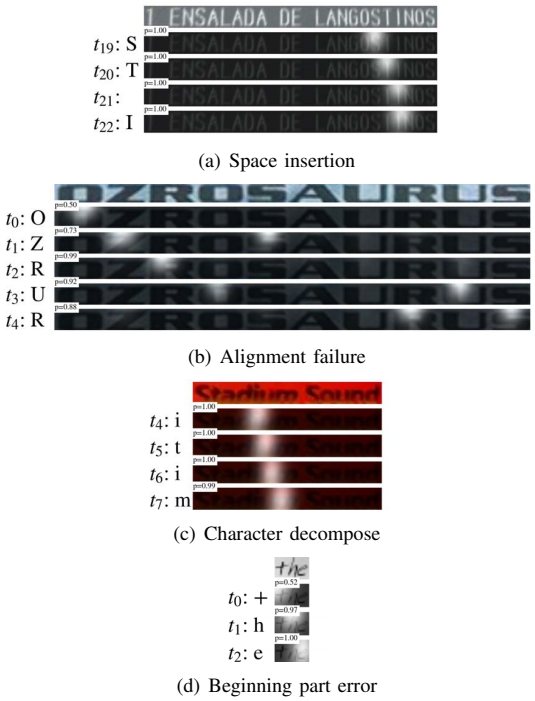
TABLE VII
COMPARISON OF ATTENTION AND CTC ON SINGLE WORD IMAGES

		G-1	S-1	11Seg-1	IC13
WER	Attention	4.8	3.0	4.2	11.2
	CTC	3.9	3.0	3.5	12.7
SER*	Attention	5.0	3.4	6.9	10.0
	CTC	4.8	3.5	7.0	12.1

E. Experiments on Isolated Word Recognition Tasks

Besides sentence recognition tasks, we further conduct a comparative study on isolated word datasets. The results are shown in Table VII. G-1, S-1 and 11Seg-1 are single word image sets contained in G, S and 11Seg, respectively. SER* stands for *space*-free sentence error rate, where the *space* in each hypothesis is removed with the assumption that only one single word is contained in each image. Therefore, 1-SER* equals to the accuracy measured in none lexicon mode as used in [1]–[6], [28].

The attention-based system achieves an overall lower or comparable SER* than the CTC-based system, yet yields a higher WER on G-1 and 11Seg-1 sets. This observation indicates that the attention-based system yields inferior performance in recognizing *space* than other characters. It tends to emit a *space* label when the space between two adjacent characters is slightly larger than normal inter-character space, which is known as segmentation errors. We conjecture that this is resulted from the decoding mechanism of the attention model, which relies on the last predicted label and a local context vector to predict the next character label. Therefore, it is unable to utilize a global context properly in decoding stage. However, in the CTC-based system, the whole image features are leveraged by BLSTM layers when predicting labels for each frame. Consistent with observations in the literature [4], [31], the attention-based approach can achieve better accuracy than CTC-based approach on isolated word recognition scenarios. But as shown in our experiments, it performs worse when taking the segmentation errors into account in an evaluation metric.



(a) Space insertion

(b) Alignment failure

(c) Character decompose

(d) Beginning part error

Fig. 3. Attention visualization of typical errors.

F. Error Analysis of Attention-based Encoder-Decoder System

By analyzing the decoding results and visualized attention weights, we find that there are mainly four typical error patterns in the attention-based systems, which are described below in the order of decreasing severity:

- **Word or character segmentation errors:** As illustrated in Fig. 3(a), if the space between two adjacent characters (e.g., between 'T' and 'I' in 'LANGOSTINOS') is slightly larger than a normal inter-character space, the attention model tends to insert an extra *space* label within the word. On the other hand, it will sometimes decomposes a wider character into two (e.g, 'u' is decoded as 'ti'), as illustrated in Fig. 3(c). These two kinds of errors might be resulted from the prediction mechanism of the attention model which is unable to utilize the full context or neighborhood context effectively when predicting the next character.
- **Beginning part errors:** As illustrated in Fig. 3(d), the attention model is unable to predict the beginning character reliably, especially when the beginning character is ambiguous (e.g., 't' is decoded as '+'). This is because there is little context to be leveraged when predicting the first character since the decoding process is performed from left to right in an auto-regressive way.
- **Alignment errors:** this is a typical and well-known issue of the attention model in NLP field [18], [19]. The attention model may sometimes miss a segment or decode the same segment many times. As illustrated in the second and fourth rows in Fig. 3(b), the attention model is confused and unable to find the next character correctly when the same character 'o' is observed at more than two positions. The position embedding [32] or the coverage model can alleviate this problem to a certain extent.

V. Summary and Future Work

We have conducted a comparative study of different encoder, attention and decoder modules in an attention-based encoder-decoder approach and compared its performance with a CNN-DBLSTM-CTC based system on large-scale real-world scene text sentence recognition tasks. Our experimental results show that the attention-based approach can achieve a better result than the CTC-based approach on isolated word recognition tasks when decoding with a greedy search strategy, which is largely consistent with the observations made in the literature. However, the attention-based approach performs worse than the CTC-based approach on scene text sentence recognition scenarios and/or when an explicit language model is used in decoding.

We further analyze and illustrate some common error patterns of the attention-based systems. Among these errors, the issue of segmentation errors is the most serious one. As future works, we will try to alleviate this problem from three aspects: 1) Learning a seq2seq module to map a character sequence to a word sequence; 2) Decoding with a smarter search strategy or a more powerful language model; 3) Making better use of context information in attention mechanism.

References

[1] C.-Y. Lee and S. Osindero, "Recursive recurrent nets with attention modeling for ocr in the wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2231–2239.

[2] B. Shi, X. Wang, P. Lyu, C. Yao, and X. Bai, "Robust scene text recognition with automatic rectification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4168–4176.

[3] Z. Cheng, F. Bai, Y. Xu, G. Zheng, S. Pu, and S. Zhou, "Focusing attention: Towards accurate text recognition in natural images," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 5086–5094.

[4] H. Li, P. Wang, C. Shen, and G. Zhang, "Show, attend and read: A simple and strong baseline for irregular text recognition," *arXiv preprint arXiv:1811.00751*, 2018.

[5] 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," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 11, pp. 2298–2304, 2017.

[6] S. Fang, H. Xie, Z.-J. Zha, N. Sun, J. Tan, and Y. Zhang, "Attention and language ensemble for scene text recognition with convolutional sequence modeling," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 248–256.

[7] M. Yousef, K. F. Hussain, and U. S. Mohammed, "Accurate, data-efficient, unconstrained text recognition with convolutional neural networks," *arXiv preprint arXiv:1812.11894*, 2018.

[8] Y. Zhu, Z. Xie, L. Jin, X. Chen, Y. Huang, and M. Zhang, "Scut-ept: New dataset and benchmark for offline chinese text recognition in examination paper," *IEEE Access*, vol. 7, pp. 370–382, 2019.

[9] H. Ding, K. Chen, Y. Yuan, M. Cai, L. Sun, S. Liang, and Q. Huo, "A compact cnn-dblstm based character model for offline handwriting recognition with tucker decomposition," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 507–512.

[10] W. Hu, M. Cai, K. Chen, H. Ding, L. Sun, S. Liang, X. Mo, and Q. Huo, "Sequence discriminative training for offline handwriting recognition by an interpolated ctc and lattice-free mmi objective function," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 61–66.

[11] J. Zhang, J. Du, S. Zhang, D. Liu, Y. Hu, J. Hu, S. Wei, and L. Dai, "Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition," *Pattern Recognition*, vol. 71, pp. 196–206, 2017.

[12] F. Sheng, Z. Chen, and B. Xu, "Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition," *arXiv preprint arXiv:1806.00926*, 2018.

[13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[14] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 116–131.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 770–778.

[16] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations (ICLR)*, 2015.

[17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems 28*, 2015, pp. 577–585.

[18] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1412–1421.

[19] Z. Tu, Z. Lu, Y. Liu, X. Liu, and H. Li, "Modeling coverage for neural machine translation," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1, 2016, pp. 76–85.

[20] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Trajectory-based radical analysis network for online handwritten chinese character recognition," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 3681–3686.

[21] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.

[22] C. Gulcehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio, "On using monolingual corpora in neural machine translation," *arXiv preprint arXiv:1503.03535*, 2015.

[23] A. Sriram, H. Jun, S. Satheesh, and A. Coates, "Cold fusion: Training seq2seq models together with language models," *arXiv preprint arXiv:1708.06426*, 2017.

[24] I. Krasin, T. Duerig, N. Alldrin, A. Veit, S. Abu-El-Haija, S. Belongie, D. Cai, Z. Feng, V. Ferrari, V. Gomes, A. Gupta, D. Narayanan, C. Sun, G. Chechik, and K. Murphy, "Openimages: A public dataset for large-scale multi-label and multi-class image classification." *Dataset available from https://github.com/openimages*, 2016.

[25] 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 *Document Analysis and Recognition (ICDAR), 2013 12th International Conference on*. IEEE, 2013, pp. 1484–1493.

[26] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, "Accurate, large minibatch sgd: Training imagenet in 1 hour," *arXiv preprint arXiv:1706.02677*, 2017.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *international Conference on Learning Representations (ICLR)*, 2015.

[28] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2018, pp. 71–79.

[29] H. Ding, K. Chen, W. Hu, M. Cai, and Q. Huo, "Building compact cnn-dblstm based character models for handwriting recognition and ocr by teacher-student learning," in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*. IEEE, 2018, pp. 139–144.

[30] M. Cai, W. Hu, K. Chen, L. Sun, S. Liang, X. Mo, and Q. Huo, "An open vocabulary ocr system with hybrid word-subword language models," in *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, vol. 1. IEEE, 2017, pp. 519–524.

[31] S. K. Ghosh, E. Valveny, and A. D. Bagdanov, "Visual attention models for scene text recognition," in *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1. IEEE, 2017, pp. 943–948.

[32] 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.