Transformers for Historical Handwritten Text Recognition

Killian Barrere, Yann Soullard, Aurélie Lemaitre, Bertrand Coüasnon

Transformers for Historical Handwritten Text Recognition

Killian Barrere
PhD Student
killian.barrere@irisa.fr

Yann Soullard
PhD Supervisor
yann.soullard@irisa.fr

Aurélie Lemaitre
PhD Director
aurelie.lemaitre@irisa.fr

Bertrand Coüasnon
PhD Director
bertrand.couasnon@irisa.fr

Univ Rennes, CNRS, IRISA, France

This work is part of a thesis entitled
"Deep Neural Networks and Attention Mechanisms for Handwritten Text Recognition".
The thesis started on October 1st 2020 and is expected to finish by September 30th 2023.

***Abstract*—Handwritten documents are recently getting more and more publicly available, but searching efficiently information through them is difficult. Handwritten Text Recognition systems automatically transcribe documents and offer excellent solutions to make the content of handwritten documents available. Neural networks are currently the state-of-the-art approaches for this task. Recently, Transformer architectures have gained in popularity in many fields. We present the works we have done so far toward an efficient architecture using transformer layers for the field of Handwritten Text Recognition. Architectures using Transformer for Handwritten Text Recognition are presented. Our architectures aim to replace recurrent layers with transformers, while combining optical recognition and language modeling in end-to-end model. We manage to obtain state-of-the-art results on the IAM dataset with one of our architecture.**

I. Introduction

Nowadays, a considerable number of handwritten documents have been digitalized and made available to the public to ease their access. However, searching information through them efficiently remains a complex task. While human transcribers could be considered to make the textual content available, this is typically a long and expensive process. This is especially true for difficult historical documents, where even an experienced transcriber could experience difficulties to decipher the textual content.

Offline Handwritten Text Recognition aims to automatically read scanned handwritten documents and output a computer-understandable text. However, they require text-line images to perform their task. Usually, a first step of document layout analysis segments the text from a page into text-line images. Handwritten text recognition systems are then used to obtain their transcripts. Finally, a post-processing step consisting of applying a Language Model is applied to correct eventual errors.

While initial models in the field of handwritten text recognition were mostly based on Hidden Markov Models, deep learning and neural networks have been showing groundbreaking improvements in the field [1]. Model based on recurrent layers have been widely used in the following years [1], [2], [3] thanks to their abilities to model sequential dependencies, paired with the well-known Connectionist Temporal Classification [1]. Convolutional layers have then been considered and added into existing architectures [4], [5].

Recent models of Handwritten Text Recognition perform very well and obtain low error rates on common datasets. However, for harder documents like historical documents, existing models often fall short to obtain low error rates. This is principally due to the inherent difficulties of historical documents caused by complex writings, various styles, deteriorated documents and old languages. These difficulties also impact skilled transcribers which usually result in a longer and costly annotation process. Therefore, historical documents typically contain a very few amount of annotated data. Such documents are challenging for existing models and remain interesting data with much room for improvement before attaining low error rates.

In the field of Natural Language Processing, Transformer models using the so-called Multi-Head Attention have been proposed and showed ground-braking results [6], [7]. Vaswani et al. introduced Multi-Head Attention, which is able to handle long-range context, while providing efficient parallelism, which is crucial in current deep learning approaches. Multi-Head Attention is, therefore, an efficient replacement to long used recurrent layers. In the years that follow, Transformer models have been employed more and more in the field of Natural Language Processing, while also showing promising results in other fields like Automatic Speech Recognition [8], Natural Image Classification [9] or even more recently in the field of Handwritten Text Recognition [10], [11].

In this work, we aim to use Transformer to recognize complex handwritten text in historical documents. As Transformer models represent an efficient alternative to recurrent neural networks, we aim to exploit their superior training capabilities to overcome the gap between the results obtained

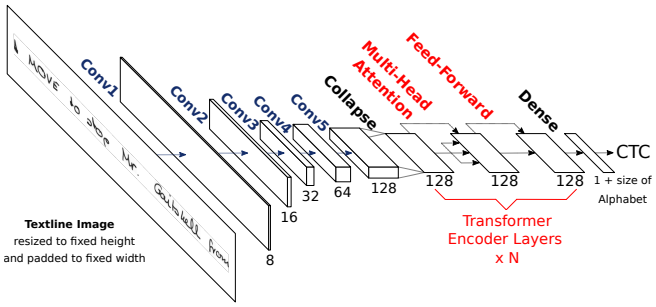

Figure 1. Our Convolutional Transformer Model.

with existing models and low error rates. Furthermore, we believe such models represent an efficient alternative to Language Models, that might be difficult to apply for historical documents, and we aim to combine both the Optical Recognition step and the Language Modeling in an end-to-end model, as achieved by Kang et al. [10].

Transformer models for Handwritten Text Recognition are introduced in this work. In Section II, we introduce Convolutional Transformer, directly derived from Convolutional Recurrent Neural Network where we replaced recurrent layers. Section III then follow and proposes a Sequence-to-Sequence Transformer where we included a Transformer Decoder to combine both Optical Recognition and Language Modeling in an end-to-end fashion. Such architectures are then evaluated on common datasets with preliminary results in Section IV.

II. Convolutional Transformers

Convolutional Recurrent Neural Networks (CRNN) are widely used for Handwritten Text Recognition [12], [13]. They are good enough to extract local and global features, while providing good information from the past and future context for text lines images by combining both convolutions and recurrent layers.

As Transformers models introduced Multi-Head Attention as an efficient alternative to recurrent layers, we propose replacing the recurrent layers in CRNN with Multi-Head Attention. However, applying transformer to 2D images is far more difficult than applying it to 1D data [9]. To alleviate this issue, we use the inherent sequentiality related to the reading order (from left to right for Latin scripts) of text-line image. This is made possible by a vertical dimension collapsing enabling our model to work on a sequence of features. Connectionist Temporal Classification (CTC) [1] is used to handle the differences in sizes between the predicted character probability and the ground truth.

Our architecture (illustrated in Figure 1) follows the general trend of CRNN architectures. It is composed of a sequence of five convolutional layers aiming to extract local features from the images. The image is quickly reduced in size with 2x2 max pooling following each of the first 3 convolution layers. Following the convolutional backbone, we collapse the vertical dimension with a convolution layer resulting in a sequence of features containing meaningful information for each column of the text-line image. We then use stacked Transformer encoder layers as an alternative to recurrent layers, that are in charge of handling long range dependencies. Each Transformer encoder layer is composed of a layer of Multi-Head Attention and a Feed-Forward Layer, with each sub-layer using residual connections. Following these stacked transformer layers, we apply a dense layer with a number of output neurons equals to the size of the character set, while including the CTC Blank character, before training it with an usual CTC loss function.

III. Seq2Seq Transformers

Following works on Convolutional Transformer, we tried to propose another model closer to the initial Transformer model by proposing a sequence-to-sequence encoder-decoder (seq2seq) architecture.

Seq2seq models for Handwritten Text Recognition takes as input text-line images as well as the sequence of what the model has already predicted. They output characters one by one. This process, therefore, enables the architecture to model the language and adapt its output character based on the characters before.

With such architecture, we aim to combine both the Optical Recognition with the Language modeling inside mutual Multi-Head Attention layers, therefore resulting in an end-to-end model. Mutual Multi-Head Attention layers combine output of the layer before it with the encoding matrix, obtained after feeding the text-line image to the encoder layers.

Our Sequence-to-Sequence Transformer model (illustrated in Figure 2) is composed of a CRNN (or a Convolutional Transformer) as the encoder, and of a stack of Transformer decoder layers as the decoder. The encoder is used to encode the text-line image in a matrix shape, which is directly obtained from the previous hidden layer. The decoder takes as input the sequence of what the model has already produced. A character-level embedding and positional encoding are applied before feeding that sequence to the decoder. However, at training time, we employ teacher forcing. We feed the target transcription to the decoder, after being shifted right to assure that the model only sees the previous characters and not the characters it should output.

Following the work done by Michael et al. [14], we find it beneficial to use a hybrid loss to train the model. Therefore, in addition to Cross Entropy Loss used to train seq2seq models, we additionally use a CTC loss function which is applied to the output of the encoder (following a dense layer).

IV. Preliminary Results

We evaluated our architecture on the IAM Dataset consisting of modern English text-line images and the READ

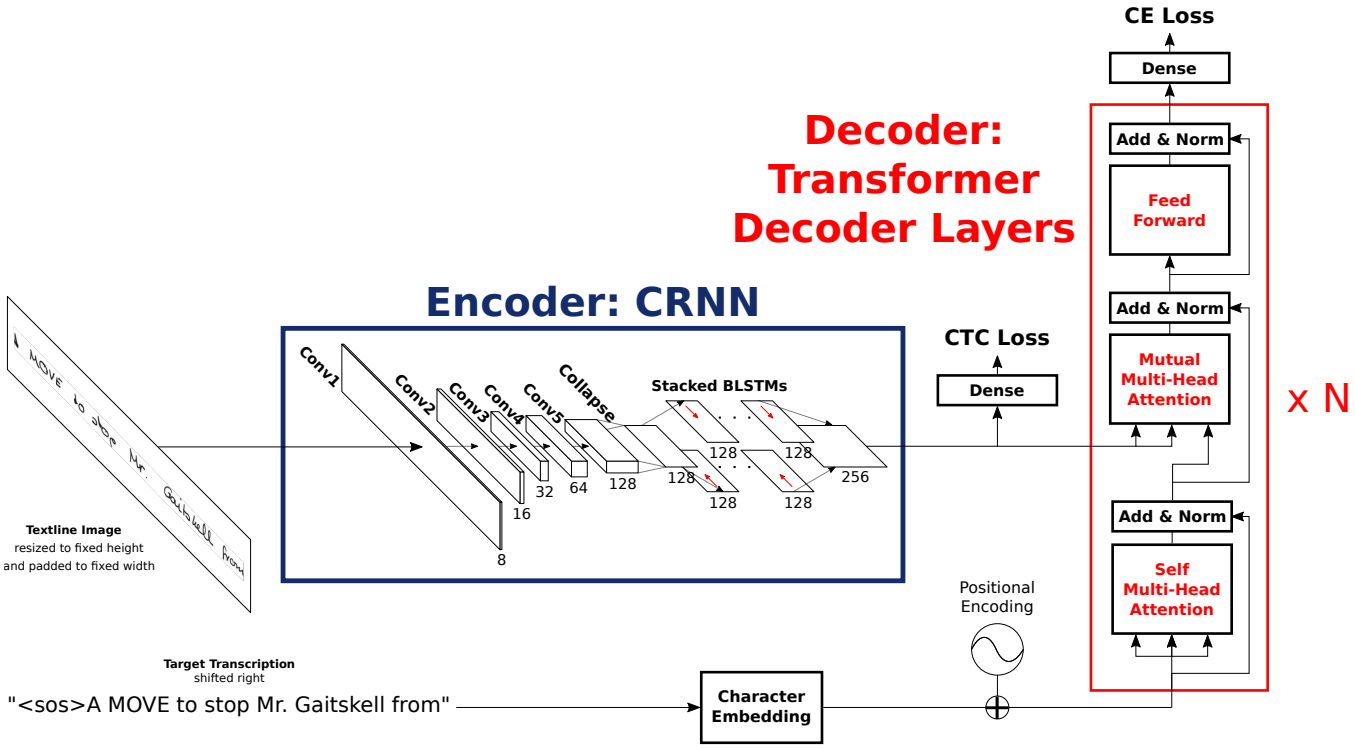

Figure 2. Our Sequence-to-Sequence Transformer Model

Table I
RESULTS ON IAM DATASET (MODERN ENGLISH).

Model Architecture	Validation		Test	
	CER	WER	CER	WER
GCRNN + additional data + LM [4]			**3.2**	**10.5**
FCN [15]	**3.3**		4.9	
Seq2Seq CRNN + LSTM [14]			4.87	
VAN [16]	**3.32**	**13.60**	4.95	16.24
Transformer [10]			7.62	24.54
Transformer + synthetic data [10]			4.67	15.45
Our Convolutional Transformer	5.99	24.17	7.42	29.09
Our Seq2Seq Transformer	4.02	16.55	5.07	21.47

Table II
RESULTS ON READ 2018 (HISTORICAL DOCUMENTS). THE DATASET CONSISTS OF 17 HISTORICAL DOCUMENTS FROM VARIOUS LANGUAGES WITH OUR CUSTOM TRAIN, VALIDATION, AND TEST SPLITS.

Model Architecture	CER	WER
Our CRNN	14.27	**39.35**
Our Convolutional Transformer	16.76	45.70
Our Seq2seq Transformer	**12.81**	41.00

2018 dataset [5], which is composed of 17 documents from various languages. For the READ dataset, we used our own custom train, validation and test split. Table I and Table II report the results for each dataset respectively. We also compared our results with the state of the art on the IAM dataset.

Regarding our Convolutional Transformer, we only managed to train small models. It obtains results under the state of the art. Despite the fact that it might be capable to possess more capabilities than typical CRNN models, we find it difficult to exploit such model. Despite that, with such architecture, we managed to obtain promising results. However, we believe there is still room for improvement.

Our seq2seq Transformer meanwhile managed to obtain state-of-the-art results on the IAM dataset without additional data. On the READ dataset, we also obtained a character error rate below what we obtain with a regular CRNN. This architecture shows very promising results, however, we find it difficult to train as it has many hyper-parameters and still have difficulties to converge efficiently. As future works, we plan to invest these problems and pursue our work toward improving further our results with this architecture.

Transformer models, nonetheless, remain a challenging architecture for historical datasets. As they require a fair amount of annotated data to obtain the best out of the architecture.

V. CONCLUSION AND FUTURE WORKS

In this work, we presented two architectures using Transformer for Handwritten Text Recognition. We use Transformer to replace recurrent layers of a CRNN, resulting in

a Convolutional Transformer model. We proposed a second architecture: seq2seq Transformer in which we included a Transformer Decoder. With this architecture, we aim at combining both Optical Recognition and Language Modeling in an end-to-end fashion. We obtain state-of-the-art results on the IAM dataset with our seq2seq transformer with a character error rate of 5.07 on the test set, without additional data. However, we find it difficult to train such models.

Ongoing works are dedicated to efficiently training such architectures by improving the training optimization, while also using more data augmentation techniques and synthetic data. This is significant as these models seem to require a substantial amount of annotated data, whereas historical documents contain few annotated data. Furthermore, we would like to work on augmentation techniques fitted especially for historical documents.

Upcoming works will be dedicated to pushing the results further, while focusing our works on historical documents.

Following that, we hope to publish our works, therefore providing a more detailed view of our models and the training procedure while making our code available to ease reproduction.

ACKNOWLEDGMENTS

This work was granted access to the HPC resources of IDRIS under the allocation 2021-AD011012550 made by GENCI.

REFERENCES

[1] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[2] T. Bluche, J. Louradour, M. Knibbe, B. Moysset, M. F. Benzeghiba, and C. Kermorvant, "The a2ia arabic handwritten text recognition system at the open hart2013 evaluation," in *2014 11th IAPR International Workshop on Document Analysis Systems*. IEEE, 2014, pp. 161–165.

[3] V. Frinken and S. Uchida, "Deep blstm neural networks for unconstrained continuous handwritten text recognition," in *13th ICDAR*. IEEE, 2015, pp. 911–915.

[4] T. Bluche and R. Messina, "Gated convolutional recurrent neural networks for multilingual handwriting recognition," in *14th IAPR ICDAR*, vol. 1. IEEE, 2017, pp. 646–651.

[5] T. Strauß, G. Leifert, R. Labahn, T. Hodel, and G. Mühlberger, "Icfhr2018 competition on automated text recognition on a read dataset," in *216th ICFHR*. IEEE, 2018, pp. 477–482.

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

[7] 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.

[8] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7829–7833.

[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[10] L. Kang, P. Riba, M. Rusiñol, A. Fornés, and M. Villegas, "Pay attention to what you read: Non-recurrent handwritten text-line recognition," *arXiv preprint arXiv:2005.13044*, 2020.

[11] S. S. Singh and S. Karayev, "Full page handwriting recognition via image to sequence extraction," *arXiv preprint arXiv:2103.06450*, 2021.

[12] J. Puigcerver, "Are multidimensional recurrent layers really necessary for handwritten text recognition?" in *14th IAPR ICDAR*, vol. 1. IEEE, 2017, pp. 67–72.

[13] Y. Soullard, W. Swaileh, P. Tranouez, T. Paquet, and C. Chatelain, "Improving text recognition using optical and language model writer adaptation," 2019.

[14] J. Michael, R. Labahn, T. Grüning, and J. Zöllner, "Evaluating sequence-to-sequence models for handwritten text recognition," in *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 2019, pp. 1286–1293.

[15] 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.

[16] D. Coquenet, C. Chatelain, and T. Paquet, "End-to-end handwritten paragraph text recognition using a vertical attention network," *arXiv preprint arXiv:2012.03868*, 2020.