ssessing Robustness of the TrOCR Model in the Presence of Noise

Chandler Nielsen
Department of Statistics
University of Michigan
nn rbor, MI, United States
chandle@umich.edu

***bstract*—Document digitalization is becoming increasingly important as physical items are being uploaded to computers for enhanced security and ease of retrieval. Moreover, with the advent of large language models, it would be desirable to have large texts summarized by language models. In this article, we explore an architecture called the transformer optical character recognition (TrO R) model which is used for text recognition. The TrO R model has been demonstrated to be the state-of-the-art in text recognition when compared to previous models. We demonstrate that the TrO R model is remarkably robust in the presence of Gaussian noise, making it an especially attractive option in the event of poor scanning or data corruption.**

The GitHub repository containing all code for this work can be found here

***Index Terms*—transformers, encoder-decoders, optical character recognition**

I. Introduction

Document digitalization is the process by which physical information such as contracts, financial information, medical images, and more can be converted to a digital format such as PDF or JPEG file. Document digitalization offers the possibility of enhanced security for one's documents and ease of analysis for medical data. This value notwithstanding, there are some potential problems with the document digitalization process. In particular, it may be that data files are imperfectly read by a neural network model, which can make character recognition and the processing of medical images problematic. In this paper, we explore the application of deep learning models for optical character recognition.

The use of deep learning models for optical character recognition is not new. Indeed, this line of work dates back to at least 1989 when Waibel et al used time-delayed neural networks for phoneme identification [5]. In 1998, Lecun and Bengio published a famous paper detailing the use of neural networks, particularly convolutional neural nets, for optical character recognition [2]. Text recognition made significant improvements with the discovery of the transformer [4]. This work relies heavily on pre-trained transformer models both for understanding images and for generating text.

In particular, we hope to analyze the performance of a fairly recent optical character recognition transformer model. We assess the robustness of the encoder-decoder model presented by Li et al [3] under the presence of Gaussian noise. This model is known as the Transformer Optical Character Recognition (hereafter **TrO R**) model. This model applies a transformer architecture for both reading in and understanding the image containing printed or hand-written text and for generating text based on this character recognition. s described in Li et al, the TrOCR model performs better than all other optical character recognition models on printed and handwritten text recognition tasks. For this project, we will focus our attention on handwritten text recognition tasks.

We began by determining the model accuracy on text from the I M handwritten database [1]. This database includes handwritten English words that can be used to train models and assess their accuracy. We restrict ourselves only to the words in the dataset. We determine the accuracy (correct identifications over total identifications) of the TrOCR model on the basic I M handwritten database and compare to the results found in [3]. Then, we steadily increase the noise on the handwritten images and determine the model accuracy on the same dataset for different amounts of noise. This will give us an indication of the performance of state-of-the-art models on data that have been corrupted.

II. Method

We begin by discussing the model that we will apply to this dataset, followed by a description of the dataset and the process by which we will make the data noisy.

. *The TrOCR Model*

TrOCR was introduced by Li et al [3] as a competitor to text recognition models employing convolutional neural networks (CNN) for the encoder for image understanding and a recurrent neural network (RNN) for text generation. s of the publishing of this work, most models were based on self-attention with a CNN backbone to understand text images. Moreover, the RNNs were coupled with an external language model to improve accuracy. The foregoing paper demonstrates that using a pre-trained image and text transformer performs better than the state-of-the-art models without the need for an external language model. Thus, the TrOCR is simple but effective and very easy to implement. In particular, we use the $\text{TrOCR}_{\text{B SE}}$ model described in [3] (see Table 3), involving 334 million parameters. The model used for this work was

downloaded from the online I community HuggingFace here. This model was fine-tuned on the I M handwriting database, described next.

B. The I M Dataset

The I M handwritten database [1] is a database of handwritten characters maintained by the University of Bern by the Research Group on Computer Vision and rtificial Intelligence. It contains 5,685 labeled English sentences and 115,320 English words labeled in the dataset. The sentences and words of the dataset originate from 500 different authors. s described on the I M database website, the words were extracted from pages of scanned text using an automatic segmentation scheme. The words scanned were then verified manually. Each picture is in a portable networks graphics (PNG) format. The individual pictures consist of texts of the size L_0 pixels by W_0 pixels. On the other hand, the TrOCR model requires 3 color channels for each pixel, so it was necessary to convert each image to RGB using the Pillow library in Python before feeding those pictures into the transformer model. Ultimately, the input data to the model were in the space ^{L} ^{W} $^{r\,g\,b)}$, where each of r, g, b lies in $[0, 255]$. On the other hand, the output of the model consisted of English words. These could be compared with the labels provided in the IM dataset.

Some of the text in this dataset is already very messy. Indeed, as Figure 1 below illustrates, it can be difficult for human beings and machines to read the basic text even before the addition of noise.

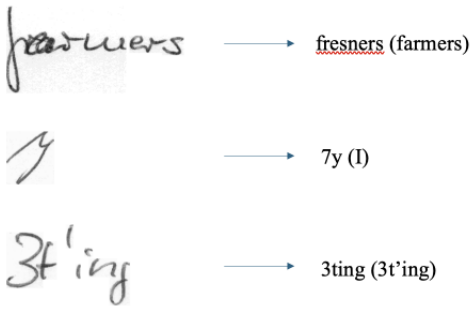

Fig. 1. Pictures to the words mapped by the model. The words after the arrows are the outputs of the TrOCR model; the words in parentheses are the correct mappings.

s we explain in the section Results below, this potential difficulty in determining the correct token was the impetus for introducing randomness to eliminate any bias that may have resulted from the model encountering tokens that are already difficult to read due to sloppy handwriting. Next, we proceed to the means by which we added noise to the text images read by the TrOCR model.

C. dding Noise

To determine the robustness of the TrOCR model to corruption of the text images it encounters, we added progressively more noise to each of the images read by the model. For each image, we create a L_0 W_0 3 array of Gaussian random variables $X \sim \mathcal{N}\ 0,\ ^{2})$. We add this noise to each basic image for different values of . For this paper, we explored the addition of Gaussian noise when $\in \{0, 80, 100, 316.23\}$. n illustration of this process is provided in Figure 2 below.

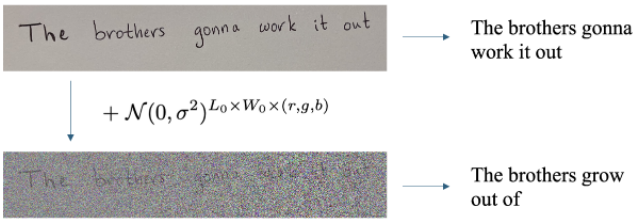

Fig. 2. Pictures to the words mapped by the model. The words after the arrows are the outputs of the TrOCR model; the words in parentheses are the correct mappings.

Figure 2 demonstrates the extent to which the model struggles after the addition of noise. We sought to determine how robust this state-of-the-art model is to the addition of noise, and whether there exists some critical after which the model performance decreases drastically. Note that the prediction made by the model in the presence of noise will depend on the iteration, partly as a consequence of using a new random three-dimensional array for each new image.

With all of this in mind, we proceed to our results.

III. Results

We begin this section by describing our data analysis pipeline and the metrics used to evaluate model performance under varying levels of noise. Next, we present the results. Finally, we interpret these results.

. *The Data nalysis Pipeline*

Due to the inherent noise present in different handwriting styles, we decided to use randomness to mitigate the effects of the relative legibility for different authors represented in the dataset. Namely, we analyzed the 5685 English sentences in the dataset. We first excluded from selection those sentences that were used to fine-train the model. Then, we sample 1000 sentences without replacement. These 1000 sentences were used to assess model performance for each value of $\in \{0, 80, 100, 316.23\}$. Each sentence was separated into tokens by English words and punctuation. For each one of these sentences, the model's accuracy was deemed 'correct' if the model correctly output at least 70% of the tokens in the input text image. The performance of the model under different levels are noise are presented in the next subsection.

B. Model Performance

Each run/row in Table 1 corresponds to the same set of 1000 sentences being analyzed. Note that there is a considerable amount of variation between each set of 1000 sentences for each . This justifies the use of different sets of sentences; the legibility varies wildly for different authors.

T BLE I
CCUR CY OF MODEL UNDER DIFFERENT LEVELS OF G USSI N NOISE

Run Number	No noise, =	= 8	= 1	= 316.23
Run 1	0.870	0.822	0.780	0.163
Run 2	0.874	0.804	0.764	0.160
Run 3	0.833	0.816	0.779	0.186
Run 4	0.896	0.832	0.785	0.176

C. Interpretations

Note that the TrOCR model performs significantly better than the PaddleOCR model that was considered in the project proposal. The latter model is built on a CNN coupled with a RNN for text generation; this is in line with the conclusiosn reached by Li et al in the paper motivating this work.

nother interesting feature of these results is how drastically the accuracy decreases once the noise reaches a certain threshold. Because of the relatively high computational cost of analyzing each set of 1000 sentences, there was not time to conduct this work for values of for a finer grid of values. It would be interesting to produce a curve to determine the performance of the model for different values of and determine if there exists a level of that causes a drastic decrease in model performance.

Finally, we feel that it is worthy of note that an image with text and $= 316.23$ (corresponding to a variance of 100000) is still mostly legible to a human reader (Figure 2 may not illustrate this point because it has been reduced in size). It would be interesting to attempt to compare the performance of human beings in read these text samples and comparing their performance to the performance of TrOCR.

IV. CONCLUSIONS ND FUTURE DIRECTIONS

The transformer based optical character recognition model is the state of the art; Lit et al have already demonstrated its capabilities relative to optical character recognition models based on CNNs and RNNs. In this article, we sought to determine the extent to which this transformer model is sensitive to noise. Even in the presence of a fair amount of Gaussian noise ($= 80$), the model continues to perform even on datasets with fairly poor handwriting. It is only until a large level of Gaussian noise is reached ($^2 = 100000$) that the model's text recognition begins to fail. This project can be extended in a number of ways. Due to lack of time and high computational costs, we were not able to perform this procedure on the full I M dataset. It would be interesting, if given these resources, to see if the transformer model would still be regarded as the state-of-the-art in the presence of increasing amounts of Gaussian noise. Given that a relatively large amount of noise is required before the model begins to fail, we suspect that the answer to this would be yes. Finally, it would be interesting to see how well the TrOCR model performs relative to humans transcribing text. This model should definitely be considered if one must scan physical documents and have its contents read quickly. n optical character recognition model like TrOCR coupled with a large language model could even provide summaries of large physical documents, which would be useful for a number of fields like law, medicine, and business. We look forward to advancements toward these ends in the future.

REFERENCES

[1] H. Bunke and U.-V. Marti. "The I M-database: an English sentence database for offline handwriting recognition". In: *International Journal on Document nalysis and Recognition* 5 (2002). URL: https://link.springer.com/article/10.1007/s100320200071.

[2] Y. LeCun et al. "Gradient-Based Learning pplied to Document Recognition". In: *Proceedings of the IEEE* 86 (1998). URL: http://vision.stanford.edu/cs598_spring07/papers/Lecun98.pdf.

[3] Minghao Li et al. "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models". In: *CoRR* abs/2109.10282 (2021). arXiv: 2109.10282. URL: https://arxiv.org/abs/2109.10282.

[4] shish Vaswani et al. " ttention Is ll You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[5] lex Waibel et al. "Phoneme Recognition Using Time-Delay Neural Networks". In: *IEEE Transactions on coustics, Speech, and Signal Processing* 37.3 (1989). URL: https://www.inf.ufrgs.br/~engel/data/media/file/cmp121/waibel89_TDNN.pdf.