

Assessing Robustness of the TrOCR Model in the Presence of Noise

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

Abstract—This document is a model and instructions for L^AT_EX. This and the IEEEtran.cls file define the components of your paper [title, text, heads, etc.]. *CRITIC L: Do Not Use Symbols, Special Characters, Footnotes, or Math in Paper Title or Abstract.

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 **TrOCR**) 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. As 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.

A. 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. As 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 TrOCR_{BASE} model described in [3] (see Table 3), involving 334 million parameters. The model used for this work was downloaded from the online PyTorch 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 Artificial 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.

As described on the IM 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 $L \times W \times (r, g, b)$, 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.

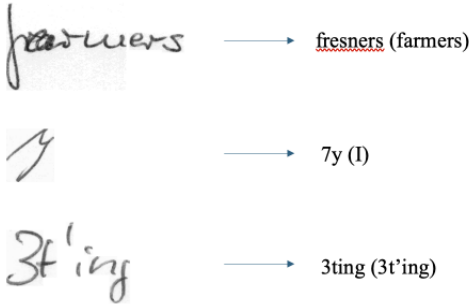


Fig. 1. Pictures to the words mapped by the model. The words after the arrows are the outputs of the TrOCR model; the words in parentheses are the correct mappings.

As 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. Adding 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 $L_0 \times W_0 \times 3$ array of Gaussian random variables $X \sim \mathcal{N}(0, \sigma^2)$. We add this noise to each basic image for different values of σ . For this paper, we explored the addition of Gaussian noise when $\sigma \in \{0, 80, 100, 316.23\}$.

An illustration of this process is provided in Figure 2 below.

Figure 2 demonstrates the extent to which the model struggles after the addition of noise. We sought to determine how

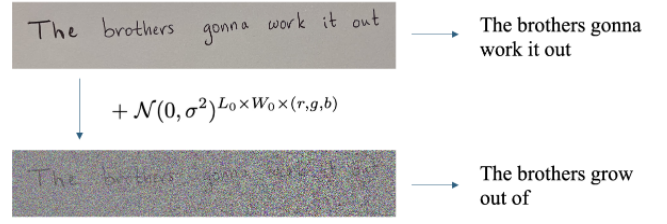


Fig. 2. Pictures to the words mapped by the model. The words after the arrows are the outputs of the TrOCR model; the words in parentheses are the correct mappings.

robust this state-of-the-art model is to the addition of noise, and whether there exists some critical σ 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.

A. The Data Analysis 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 $\sigma \in \{0, 80, 100, 316.23\}$. 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 of noise are presented in the next subsection.

B. Model Performance

TABLE I
ACCURACY OF MODEL UNDER DIFFERENT LEVELS OF GAUSSIAN NOISE

Run Number	No noise, $\sigma = 0$	$\sigma = 80$	$\sigma = 100$	$\sigma = 316.23$
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

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 level of noise σ , the accuracy on

C. Interpretations

IV. CONCLUSIONS AND FUTURE DIRECTIONS

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 Analysis and Recognition* 5 (2002). URL: <https://link.springer.com/article/10.1007/s100320200071>.
- [2] Y. LeCun et al. “Gradient-Based Learning applied 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] Ashish Vaswani et al. “Attention Is All You Need”. In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: <http://arxiv.org/abs/1706.03762>.
- [5] Alex Waibel et al. “Phoneme Recognition Using Time-Delay Neural Networks”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 37.3 (1989). URL: https://www.inf.ufrgs.br/~engel/data/media/file/cmp121/waibel89_TDNN.pdf.