A Comprehensive Evaluation of TrOCR with Varying Image Effects

Ray L. Zhang

Received September 27, 2024
Accepted November 01, 2024
Electronic access November 15, 2024

Accurate optical character recognition (OCR) is crucial for many applications such as digitizing handwritten documents, automating data entries, and real-time text recognition. In this field, a recently developed Transformer-based Optical Character Recognition (TrOCR) model utilizing the transformer architecture for both image comprehension and text generation, surpasses the current cutting-edge models on the printed, handwritten, and scene text recognition tasks. However, its performances regarding various image effects like font/background colors, font styles etc. have not been systematically investigated and revealed before, although certain deformations were applied in previous training datasets. Therefore, we evaluated how these image effects quantitatively influence the prediction accuracies at both character and word levels for this published pre-trained TrOCR model. We created 10 different synthetic printed datasets, each including 1000 randomized sentences with a comprehensive set of randomized images effects: blur, noise, rotation, font/background colors, font size, and font styles that can occur in scene text recognition, giving $>\sim$1.8 million possibilities. These combined image effects reduce the character recognition accuracy from 95% for one widely used handwritten dataset IAM to $\sim$82% for the current randomized synthetic printed datasets. A detailed analysis of all these image effects showed that the color effect is the most significant, which was not reported before. By using black font color and white background color as the typical setting in the original training datasets, even with all other factors still randomized, the character accuracy is increased to 97% and can be further enhanced to almost perfection, 99.7%, when other random effects are removed. Blur was found to be the second influential factor (removing it results in accuracy improved to 90%), while all other images effects mentioned above can be well tolerated and have no significant quantitative differences. Similar trends were observed at the word level. These results demonstrated that the pre-trained TrOCR model is indeed an excellent OCR model with no significant performance downgrade from text content, size, and style and free of noise and rotational deformation for printed text. Our data also suggest that with further training or fine-tuning to include the color and blur effects in the future, this model could be improved with superb performance beyond black/white text recognition fields. These novel results may facilitate future OCR developments using transformers to recognize more diverse real-world text backgrounds.

Introduction

Handwritten optical character recognition (OCR) is crucial for digitizing handwritten documents, automating data entry processes, and enhancing accessibility for visually impaired individuals. Accurate recognition systems can significantly improve the efficiency and accuracy of these processes, reducing manual labor and errors.

Research in OCR has evolved significantly over the past few decades in not only recognizing characters, but also in recognizing sentences and documents[1]. After an extensive review of the strengths, weaknesses, and challenges of different OCR approaches, we can clearly see a strong need for robust, adaptable OCR systems for applications like document digitization and real-time text recognition.

Literature Review

Existing research can be broadly categorized into two main groups: conventional machine learning (ML) methods and deep learning methods.

Early approaches relied heavily on feature extraction methods combined with machine learning classification algorithms[2–7] such as Support Vector Machine (SVM)[5], Random Forests (RF)[1–4,7], k Nearest Neighbor (kNN)[1–4,7], Decision Tree (DT)[2,3]. For example, an SVM[5] classifier was used for handwritten digit recognition[6] after sampling regions of local features. The feature extraction process is critical to identify key features to distinguish different classes correctly, which can be done using either statistical or structural techniques[7]. The statistical feature extraction is based on pixel distribution in an image, such as histograms, zoning, and moments. In contrast, the structural feature extraction technique utilizes the geometric characteristics of a character, e.g. loops, intersections, and number of endpoints[7]. These methods focus on simple character recognition[7] and are generally less computationally intensive but may struggle with complex datasets.

With advancements in better-performing deep learning methods and extensive use of GPUs, artificial neural network models such as Convolutional Neural Networks (CNNs) and Recurrent

Neural Networks (RNNs) have been widely used in OCR tasks due to their ability to automatically extract high dimensional features and handle complex data [1,7–9]. What is the pipeline of CNN+RNN models for OCR tasks? In these models, CNN is used for extracting visual features to understand text images and has been widely employed for classification and recognition of many different languages [10–15]. Then, the RNN uses these features to predict the text and can be improved with other methods such as Connectionist Temporal Classification (CTC), multidimensional layers [16–18].

More recently, transformer models [19–21], which have shown great success in natural language processing (NLP) but are newer in the field of computer vision, are being explored with promising results due to their ability to capture long-range dependencies in data. These models show great potential in OCR tasks. In this field, the recently developed Transformer-based Optical Character Recognition (TrOCR) model [21] is advanced in that it leverages the architecture of transformers, originally designed for NLP tasks, and borrows the concept of vision encoder, which have demonstrated strong performance in computer vision tasks, to adapt it for the domain of optical character recognition. Compared to a number of other OCR models, TrOCR's word prediction accuracy (96.4%) is among the highest when using the SROIE (Scanned Receipts OCR and Information Extraction) dataset [21] e.g. the corresponding accuracies for other state-of-the-art OCR models Tesseract and CRNN are only 57.5% and 28.7% respectively. In addition, its character error rate of $\sim$3% is among the lowest when using IAM (A dataset of handwritten English text) [22]. The study showcases TrOCR's ability to process complex documents and handwritten texts with high precision, highlighting its potential for real-world OCR applications. Its popularity and widespread use in both academic research and real-world applications highlight its effectiveness and reliability. In addition, the model's transformer-based design aligns well with the current trend of using attention mechanisms to enhance OCR tasks, making it an advanced option for a thorough comparison of text recognition methods.

Our Contribution

However, although the original authors of the state-of-the-art TrOCR models demonstrated how TrOCR effectively outperformed traditional OCR methods and other deep learning approaches including CNNs and RNNs [21], there is no such research about the transformer-based OCR models' performance on different visual factors or image effects. To fill in the research gap, the primary focus of this research is on evaluating and contrasting the accuracies at both character and word levels from using different image effects. In particular, this paper quantitatively and systematically investigates the text recognition performances of this current state-of-the-art model regarding various image effects that have not been reported before: 1) we evaluated the effect of font and background colors on prediction accuracies, which seems to have not been specifically studied before in pre-training and fine-tuning of TrOCR models. But font and background colors are important factors to evaluate since contemporary real-world text environments such as street signs, event fliers, advertisement materials, include colors; 2) although different kinds of synthetic and handwritten datasets were used in the training of TrOCR models [21], the quantitative effects of font sizes and font styles on prediction accuracies are unknown. As these are intrinsic properties of texts and whether they have big influence on OCR accuracy is important to know for model development, we also studied their effects; 3) even though a few data augmentation types have included image deformations like blur, noise, and rotation in TrOCR model training [21], their specific quantitative influences on prediction accuracy remain to be elucidated in this work. These image effects also frequently occur in real-world scenarios, so having a good understanding of their impacts on TrOCR's prediction accuracy is useful; 4) given that each data has only one of the image transformations applied in the prior development of TrOCR models [21], we investigated the prediction performance when each of the data were applied with all seven different kinds of image effects at the same time, including not only all three representative image deformations of blur, noise and rotation, but also randomized font sizes and styles, and especially the random font and background color effect not studied before. This treatment is new and much closer to real-world situations that can have multiple deformations and other image effects occurring at the same time; 5) we further compared performances of such all-random-image-effects applied datasets vs. those having only one effect fixed but with all other factors still randomized. Such comparisons were also novel and useful to reveal the more significant image effects among all studied ones. 6) We compared several batches of randomized text data with all image effects applied simultaneously to see if TrOCR's prediction accuracy is stable or not across different text content examinations. Overall, we have conducted quite a few novel comparative studies for this TrOCR model. It should be noted that the identification of specific significant factors that could influence prediction accuracy is important for future development to include them in training and/or fine-tuning for a strong all-around OCR model.

To do so, we generated a number of synthetic datasets with these image effects to address these aspects and perform a comprehensive comparison of its performances, which is essential to understand its strengths and limitations. By applying these studies to some different kinds of datasets with various image effects, this research seeks to identify how variations in text image characteristics impact the model's accuracy and overall performance to offer valuable insights into its strengths and weaknesses in handling datasets that differ in type of color, size, style, deformation, and content. This comparative analysis helped us learn more about the currently most effective OCR

technique (TrOCR) for various real-world applications to identify the most suitable specific applications and datasets, ensuring optimal performance and resource utilization.

Furthermore, this research structured the materials to highlight comparative performance metrics, detailed error analysis, and the practical implications of the findings. By doing so, it provided readers with a clear understanding of which image effects could significantly affect OCR performances and how the future developments to include such effects in training may be optimized for better performance across more diverse datasets in real-world situations. This approach not only contributes to the academic conversation but also guides future research and development in handwritten character recognition technologies.

Methods

TrOCR

As shown in Fig. 1 and described below, the core structure of TrOCR comprises an encoder-decoder framework, where the encoder processes the visual features extracted from the input handwritten text images, and the decoder generates the corresponding text sequences as output.

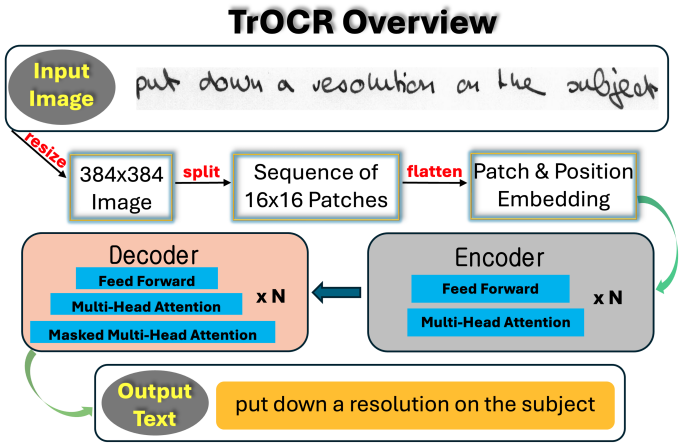

Fig. 1 Overview of the TrOCR architecture.

1. **Encoder:** The architecture of the encoder closely mirrors that of vision transformers[23], which have been successfully applied to similar tasks in computer vision. So, the encoder does not use CNN as a feature extractor, serving as a preprocessing step to distill detailed features from input images. It first resizes the input image to a fixed size of 384x384, which is then split into a sequence of 16x16 patches as input for image transformer. These patches are subsequently rendered into vectors (also known as patch embedding). The encoder also uses positional encodings to retain spatial information about the input image[24]. The patch embeddings and two special tokens (one for the whole image classification and the other for the distillation in input sequence) can be used to produce learnable one-dimensional position embedding based on their absolute positions. These extracted features and positional encodings are passed through multiple vision transformer encoder layers, including feed-forward and multi-head attention modules, which serve as the central component of the encoder and capture long-range dependencies by enabling the model to concentrate on various regions of the image at the same time, computing attention scores between features at various locations to capture diverse relationships and contextual information. This process helps the model understand complex interactions within the image. Unlike the CNN models for feature extraction, the transformer architectures do not have image-oriented inductive bias. Furthermore, treating the image as a sequence of patches allows models to more effectively focus on the entire image or specific patches.

2. **Decoder:** The TrOCR decoder converts the encoded visual features into text[21]. It operates similarly to traditional sequence-to-sequence models used in language translation, which use encoded image features for input, and the predicted text sequence as output. TrOCR utilized the original transformer decoder and has a pile of identical components. However, the decoder distributes different attention on the encoder output by adding the "encoder-decoder attention" between the multi-head self-attention and feedforward network. This process involves attending to relevant parts of the encoded features at each step, ensuring that the predicted text accurately reflects the input image. The decoder in TrOCR is responsible for transforming the encoded visual features into text sequences. After receiving the encoded image features, the decoder generates text by predicting one word at a time, using an autoregressive approach. The decoder also utilizes positional encodings to maintain the correct order of the text and relies on mechanisms like beam search during inference to enhance the quality and accuracy of the generated text sequences. In this module, both keys and values are derived from the encoder output while the queries are generated from the decoder input. Furthermore, the decoder adjusts the attention masking in the self-attention mechanism to control access to information during training compared to prediction[25].

3. **Model Initialization:**

    a. Encoder initialization: There are three levels of TrOCR models developed[21]: small, base, and large, which utilize either DeiT[26] or BEiT[27] models for encoder initialization. The DeiT[26] model was developed with different hyperparameters and data aug-

mentation to make it data-efficient and contains a robust image classification system that is condensed into a token in the starting embedding. These features result in a competitive performance when compared to CNN models. The BEiT[27] model adopts the Masked Image Modeling to pre-train the image transformer with visual tokens.

b. Decoder initialization: Both RoBERTa[28] and MiniLM[29] models were used in decoder initialization. RoBERTa[28] removed the next sentence prediction function and adapted the masking patten of the Masked Language Model in real time in order to thoughtfully manages the impact of numerous crucial hyperparameters and size of the training data. MiniLM[29] is a compact version of the large pre-trained transformer system yet maintains 99% of its performance. Aiming to circumvent the soft prediction probabilities of masked language model outputs or intermediate representations from the teacher models to facilitate the training of the student model during initial stages, it was trained by extracting the self-attention mechanism from the last transformer layer of the teacher models and incorporating a teacher assistant in early stage.

4. **Pretraining and Fine-tuning:** The TrOCR models[21] were pre-trained in two stages with different datasets. For the first stage, it utilized hundreds of millions of images with printed text lines. For the second stage, two datasets for both printed and handwritten tasks were used with millions of textline images in each case. In the second stage, distinct models were pre-trained on task-oriented datasets, and all started from the models developed in the first stage. In the original paper of TrOCR models[21], the pre-trained models underwent additional fine-tuning for the subsequent text recognition activities.

5. **Data Augmentation:** To enhance the diversity of pre-trained and fine-tuned data, the TrOCR models[21] applied data augmentations using several image transformations including random rotation ranging from -10 to 10 degrees, image blurring, resizing, erosion, and underlining. Each of these image transformations was selected with equal possibility. Other randomized data augmentations using inversion, curving, blur, noise, distortion, rotation etc were also included in its study of scene text datasets[21].

Clearly, TrOCR benefits from extensive pre-training on large-scale datasets, including both real and synthetic data, which enables it to generalize well across different handwriting styles and languages. This pretraining helps the model learn robust feature representations that are transferable to a variety of OCR tasks. The model is further fine-tuned on specific datasets, such as IAM[22], to enhance its accuracy in recognizing handwritten text. Additionally, TrOCR can be fine-tuned on other specific datasets, allowing it to adapt to various domains or languages, further improving its performance in specialized OCR applications. Three levels of TrOCR models have been developed[21]: small, base, and large, which contain total parameters of 62, 334, and 558 million.

There are several advantages of the TrOCR models[21]: 1) it does not need external language models because it builds on both pre-trained image and text transformer models; 2) it is easy to implement and maintain as it does not depend on any CNNs or any image-specific bias; 3) it does not involve complex pre/post-processing steps; 4) it can be easily extended for multilingual text recognition with adjustments of multilingual pre-trained models in the decoder side and expansion of corresponding dictionary.

Nevertheless, this model also has several weaknesses. For instance, it only works on single line sentences, which was recently extended to deal with full-page scanned receipt images[30]. As another example, since it does not contain a post-processing step and was trained with contemporary data, it may have large errors for historic hand-written text datasets[31] and mixed mode scene text test sets[21], which were recently remedied by combining it with a language model corrector[31] and by embedding a DoRA (Weight-Decomposed Low-Rank Adaptation) encoder and LoRA (Low-Rank Adapatation) decoder[32] respectively to reduce the prediction errors.

In addition, there is no specific error analysis due to different kinds of image effects, which were studied in this work.

Data Generation and Experiment

Since this study's experimental research design is aimed at comparing the performance of the recently developed advanced deep learning model TrOCR[21] on different datasets with varying image characteristics, we employed this model using the code developed by original authors (https://github.com/microsoft/unilm/tree/master/trocr) with modifications to work with the different datasets studied in this paper. The transformers package is from Hugging Face. In this work, as our focus is to evaluate the performance of the previously developed state-of-the-art OCR model TrOCR21 on various image effects, we used the published base-level pre-trained TrOCR model with no further fine-tuning to compare these effects on the same footing. This base level's accuracy data are just $\sim$1% lower than the large level ones for almost all datasets evaluated by the original developers[21]. Our evaluation used the pre-trained model without any fine-tuning to establish a baseline and assess its general robustness across various image effects, such as blur, noise, rotation, and color variations. By using the pre-trained model as is, we could better understand which image effects the model handles well and which it struggles with, without

the influence of task-specific fine-tuning. This approach provided a clearer picture of the model's inherent strengths and weaknesses when faced with distortions not explicitly addressed during training. Isolating these weaknesses in the pre-trained model helped us identify areas where future improvements could be most beneficial.

We used Python script and PyTorch[33] for model evaluation, accuracy measurement, and error analysis. The commonly used Python libraries such as NumPy, Pandas, Random, Faker, Evaluate and PIL were utilized for statistical analysis, text generation, and image processing. The Microsoft Excel software was used for data visualization and calculations of mean and standard deviations. The Google Colab served as the primary web-based interactive computing environment to employ the different tools and analysis.

The data were first collected using the IAM[22] dataset, which is one of the most popular datasets in the field of handwritten character recognition[21,32], as a validation of our code and a benchmark for comparison with synthetic data recognition. This is a dataset of handwritten English text, offering a rich variety of handwriting styles and challenges for recognition models. It is composed of 82,227 words from 400 different writers.

Additionally, synthetic data was generated using a custom procedure designed to simulate diverse text scenarios.

As shown in Table 1, we conducted a comprehensive evaluation of how prediction performances are affected by the seven different types of factors. Overall, the different random image effects when combined offer a large variety of test examples to assess the prediction capabilities of this TrOCR model on synthetic texts: $>\sim$1.8 million image effect possibilities (not including the random text size), 100 different color effects x 10 different fonts x 21 font sizes x 2 (with and without blur) x 2 (with and without noise) x 21 rotational angles (using the integers to estimate the lower ends, as actual angles are floating points and thus have much large angle ranges).

For the above purpose, we run predictions of 10 different datasets (each dataset contains 1000 text images of various effects) called Cases 1-10 in Table 2 using the TrOCR base pre-trained models (which contains 334 million parameters):

a) Cases 1, 9 and 10: All seven factors mentioned above are randomized.

b) Case 2: Texts are the same with Case 1, but all other six effects are fixed at black/white for font/background colors, for 30 font size, Arial font, no blur, no noise, and no rotation.

c) Cases 3-8: Texts are the same with Case 1, and each of the six effects from 2) to 7) as mentioned above is fixed at black/white for font/background colors, for 30 font size, Arial font, no blur, no noise, and no rotation, respectively, while all other effects are the same with Case 1 texts.

No.	Factor name	Factor range
1	Text	1000 sentences of random lengths from 1 to 9 words generated by Python module Faker
2	Color	10 different font colors: black, red, blue, green, purple, orange, brown, gray, cyan, magenta 10 different background colors: white, light gray, yellow, lightblue, light green, beige, lavender, light coral, peachpuff, honeydew
3	Font size	range of 20-40
4	Font style	10 different True Type Fonts: Arial, Constantia, Calibril, Gabriola, Lucida Sans Unicode, Times New Roman, Palatino Linotype, Perpetua, Verdana, Tw Cen MT
5	Blur	randomly applying Gaussian blur effect
6	Noise	randomly applying noise effect
7	Rotation	randomly applying rotational angle of the generated text within a range of -10 to 10 degrees

Table 1 Details of randomized text image factors

As shown above, these datasets were chosen for their diversity in complexity and imaging styles, allowing for a comprehensive comparison of the transformer-based state-of-the-art TrOCR method. These new synthetic datasets have been deposited in GitHub. It should be noted that in each case, the complete dataset was used as a test set in our evaluations of this model.

We used the following metrics to evaluate the performance. The accuracy of the TrOCR model when applied to different datasets is measured as the proportion of correctly recognized characters to the total number of characters in the test set as defined below along with its relationship to character error rate (CER) used in the original and subsequent TrOCR work[21,30–32]:

$$\text{Accuracy} = \frac{\text{Number of correct character predictions}}{\text{Number of total characters in a dataset}} = 100\% - \text{CER}$$

In addition, we evaluated the word level prediction accuracies as follows, which were also used in the original and subsequent TrOCR work[21,30,32]:

$$\text{Precision} = \frac{\text{correctly matched words}}{\text{number of detected words}}$$

$$\text{Recall} = \frac{\text{correctly matched words}}{\text{number of original words}}$$

All these comparisons are case-sensitive, so even one character with a switch in upper versus lower case would be considered wrong. A wrong character prediction results in a wrong word comparison even if other characters are recognized correctly. All four above metrics were reported in Table 2 for the studied datasets, although only accuracy, precision, and recall data were discussed in detail in subsequent Results and Discussion sections.

All these evaluation accuracies were calculated using Python scripts, specifically tailored for optical character recognition tasks. The experimental procedures of this research are as follows:

1. **Dataset Acquisition:** The dataset IAM was obtained from a credible online source (Hugging Face) commonly used in the field of optical character recognition and machine learning research[21,32].

2. **Synthetic Data generation:** Synthetic data was generated using a custom procedure as described above involving various random effects to create a diverse set of text samples. This procedure included variations in font styles, sizes, colors, backgrounds, and additional effects such as noise, blur, and rotation.

3. **Model Type:** We used the published pre-trained base TrOCR model which has 334 million parameters[21].

4. **Performance Evaluation:** The model's accuracy was evaluated on the test sets of each dataset. Comparative analysis was conducted to assess how the model's accuracy varied across datasets with different characteristics. The analysis focused on identifying significant differences in model performance across the datasets and understanding how specific dataset characteristics influenced these results.

Results

Evaluation of the predictions for a standard handwritten dataset

As shown in Table 2, the predicted character accuracy for the standard handwritten IAM dataset is 95.2%, which is excellent and close to the original paper's accuracy (96.6%) for the IAM dataset[21]. The small difference can be attributed to variations in the evaluation process and experimental setup, such as image resizing, normalization, or text cropping methods, subtle differences in the random seed used during evaluation or specific configuration details in the test set. Despite these differences, the results remain close, indicating that the model's performance on the IAM dataset is consistent with the original findings. This high accuracy result indicates that this model is indeed powerful enough to generate excellent recognition performance. The word-level prediction accuracy was not reported previously. But our results show that ∼89% words have been correctly detected and identified, which are 6% lower than the character level accuracy.

Comparison of Different Image Effects on the Predictions for Synthetic Printed Datasets

Cases 1-10 results are for the synthetic data using 1000 random text images in each dataset with various image effects.

As seen from Table 2, for Case 1 with every image factor randomized and all factors are included, its character level accuracy is 84.3%, which is significantly lower than that for IAM by 11%. Its word level precision and recall values are lower than those for IAM by 15%. For Case 2, except that texts are still randomized as in Case 1, all six types of image effects listed in Table 2 are fixed to one individual value. As illustrated in Fig.2, this case has the best and almost perfect predictions with characters correctly recognized at 99.7% and words correctly predicted with 98%. From Case 3 to Case 8, one of the six image effects was fixed at the selected value as in Case 2, while the remaining five effects are still included at the same time as in Case 1. As shown in Table 2 and Fig.2, among these six cases, both the character accuracy and word precision/recall data follow this sequence: fixed color (Case 3) $>$ no blur (Case 6) $>$ fixed font style (Case 5) $\sim$ fixed font size (Case 4) $\sim$ no noise (Case 7) $\sim$ no rotation (Case 8). In fact, compared to the accuracy results of Case 1 with every factor included and randomized, among Cases 3-8 with one fixed effect and other factors still all included and randomized, only Case 3 (fixed color) and Case 6 (no blur) have accuracy data 5% higher than Case 1. Case 3 performance is better than Case 6. Case 3 has a character accuracy improvement of ∼13% and word precision/recall improvement of ∼7% over Case 1. The data for fixed font style (Case 5), fixed font size (Case 4), no noise (Case 7), and no rotation (Case 8) are all within 3% from Case 1 data.

Examination of Robustness of the Predictions for Synthetic Printed Datasets with all Image Effects Together

We then studied and compared Cases 9 and 10 with Case 1. These three datasets have all the examined image effects simultaneously included and randomized. As shown in Table 2 and Fig. 3, the TrOCR's performances for these three datasets are similar. In fact, standard deviations from the corresponding means of the three accuracy parameters including at both the character and word levels are just around 2%, which is significantly small.

Dataset	Random font/background color	Random font size (20-40)	Random font style	Random blur	Random noise	Random rotation	CER	Accuracy	Precision	Recall
IAM							4.8	95.2	88.8	88.8
Case 1	Yes/Yes	Yes	Yes	Yes	Yes	Yes	15.7	84.3	74.1	74.0
Case 2	Black/White	30	Arial	No	No	No	0.3	99.7	98.4	98.3
Case 3	Black/White	Yes	Yes	Yes	Yes	Yes	3.0	97.0	91.5	90.9
Case 4	Yes/Yes	30	Yes	Yes	Yes	Yes	14.6	85.4	75.3	75.9
Case 5	Yes/Yes	Yes	Arial	Yes	Yes	Yes	12.8	87.2	77.8	77.3
Case 6	Yes/Yes	Yes	Yes	No	Yes	Yes	10.0	90.0	80.6	81.7
Case 7	Yes/Yes	Yes	Yes	Yes	No	Yes	14.1	85.9	75.4	76.8
Case 8	Yes/Yes	Yes	Yes	Yes	Yes	No	17.5	82.5	74.5	78.6
Case 9	Yes/Yes	Yes	Yes	Yes	Yes	Yes	19.0	81.0	70.6	72.1
Case 10	Yes/Yes	Yes	Yes	Yes	Yes	Yes	20.0	80.0	70.9	69.0

Table 2 Prediction accuracy data (in %) of the pre-trained TrOCR base model for different datasets

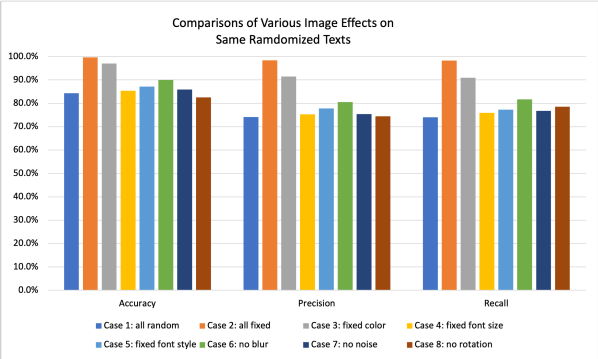

Fig. 2 Comparisons of Cases 1-8 using the same set of randomized 1000 sentences with various image effects on three groups of prediction analysis data. 1st, 2nd, and 3rd groups are for accuracy, precision, and recall, respectively, each from Case 1 to Case 8. Brief notes for Cases 1-8: Case 1 – All random; Case 2 – All fixed; Case 3 – Fixed color; Case 4 - Fixed font size; Case 5 – Fixed font style; Case 6 – No blur; Case 7 – No noise; Case 8 – No rotation.

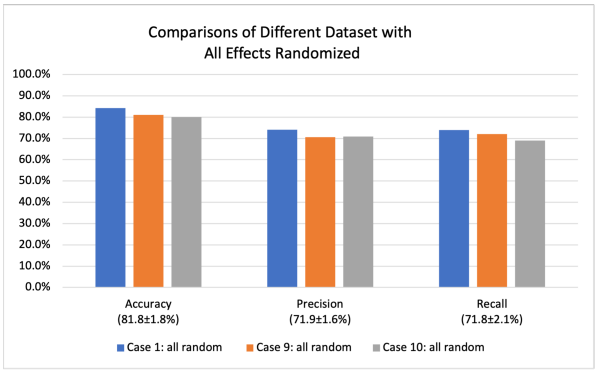

Fig. 3 Comparisons of Cases 1, 9, 10 with all effects randomized for 1000 randomized sentences on three groups of prediction analysis data. Mean±SD (standard deviation) for these three groups of data are shown.

These results indicate that the TrOCR model's performance is quite stable across different datasets with all randomized effects, which is again a strong feature of this model. Overall, the character level accuracy and word level precision/recall values are 82%, 72%, and 72%, respectively.

Correlations among Different Prediction Metrics for all Datasets

As shown in Fig.4 and Table 3, it is interesting to note that these three performance evaluation parameters are highly correlated as their corresponding linear correlation coefficient R^2 values are in the range of 0.94-0.97 and data are almost evenly distributed in their respective ranges. This feature was not reported before.

The linear regression line equations in Table 3 show that word level precision/recall results are more affected than the character level accuracy, since the slopes ($\sim$1.3) for precision vs. accuracy and recall vs. accuracy are significantly larger than one. This is reasonable as a single character recognition mistake leads to the whole word recognition error even if other characters could be correctly recognized. The similar slopes here and close to 1 slope for precision vs. recall (and as exemplified by basically the same precision/recall values, 88.8%, for the IAM datasets, and the mean precision/recall values, 72%, for the three completely randomized datasets, see Fig.3) indicate that the original words are almost 100% detected, although a few may not be recognized correctly. This almost 100% word detection rate is another nice feature of this TrOCR model, which was also found for the SROIE dataset with this model although this kind of detection rate was not directly reported before[21].

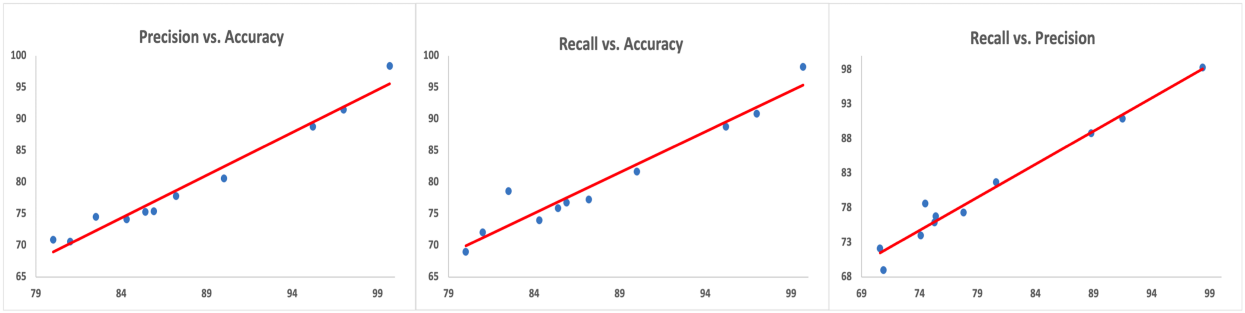

Fig. 4 Correlation diagrams among accuracy, precision, and recall data for all datasets.

Relationship	Regression Line Equation	Correlation Coefficient R^2
Precision vs. Accuracy	$y = 1.3508x - 39.086$	0.9698
Recall vs. Accuracy	$y = 1.2919x - 33.404$	0.9364
Recall vs. Precision	$y = 0.9592x + 3.7555$	0.9712

Table 3 Correlation analysis results for accuracy, precision, and recall data for all datasets

Discussion

Based on high correlations among the studied accuracy metrics here, we focused on using character-level accuracy to quantitatively investigate the various image effects studied in this work.

Cases 1, 9, and 10 were produced using the same randomization settings. Since all the factors were random and simultaneously included during the data generation process, they would have the same randomization effects applied. Since their accuracies of 84.3, 81.0, and 80.0 show a significantly small standard deviation (1.8%) from the corresponding mean of 81.8% (see Fig.3), it is good to use Case 1 dataset as the representative reference to assess the different kinds of image effects on the same randomized texts.

As such, Cases 1-8 were designed here to have the same texts for each of the 1000 images in each dataset. They provide a good basis to quantitatively compare the effect of a specific image factor on the prediction accuracy.

In general, compared to Case 1 with an accuracy of 84.3%, if Case X (X = 2-8) accuracy is similar, that means the Case X factor can be well tolerated and thus insensitive to such a type of changes. For example, Cases 4, 5, 7, and 8's accuracies of 85.4%, 87.2%, 85.9%, 82.5% are all $<3\%$ from Case 1. These four cases' specific factors different from Case 1 are the fixed font size of 30 (mean of the total font size range) for Case 3, the fixed font style of Arial (a widely used clear font) for Case 4, no noise for Case 7, and no rotation for Case 8, respectively. These results mean that constant font size and style with no noise and no rotation do not have significant improvements for recognizing the synthetic text data. Therefore, variations of these image effects do not significantly affect OCR performance of the TrOCR model. This feature indicates that the TrOCR model can recognize different font sizes, different font styles, and can tolerate the noise and rotation of the text images, which are useful for the OCR task.

Conversely, if Case X's result is greater than Case 1 by a significant amount, it means that the Case X factor has a big impact on the performance of the model. For instance, Case 6 with no blur has an accuracy of 90.0%, which is 5.7% more than Case 1. This indicates that the blur effect plays a larger role in hindering the model from recognizing the synthetic data compared to above-mentioned four image effects of font size, font style, no noise, and no rotation. This may be a result that blurring could lose some pixel details of certain features and thus decrease the prediction accuracy. However, 5.7% is still not a very large effect, which is 6.8% of the 84.3% accuracy of Case 1, i.e. still $<10\%$. This result suggests that a certain extent of blurring may be tolerated but when it affects certain texts' pixel details, it may visibly affect the prediction accuracy.

In sharp contrast, Case 2 and Case 3 have the highest two accuracies being 99.7% and 97.0% respectively. The Case 2 result indicates that when no random image effects are applied, the recognition is almost perfect, close to 100%. This showcases the strength of the model in recognizing well-structured texts. Although Case 3 has all the randomizations of image effects other than the font/background color (which is fixed at black/white as most samples in the training datasets of TrOCR), the accuracy was only slightly lower than Case 2 with no randomizations of image effects at all. This shows that when the texts are in black/white color settings as in Case 3, all other image effects including different font sizes, different font styles, random blur, random noise, and random rotation together, do not significantly downgrade the accuracy. This is likely due to the reason that the TrOCR model was trained by using texts in black/white

settings, including real life handwritings with a black pen on a white paper and computer printed receipts of black texts on a white paper[21], too. In fact, its training datasets have already included different font sizes, different font styles, and some randomization of blur, noise, and rotation effects, and thus it is not surprising to see that this model's prediction can tolerate these image effects as long as the color settings are black font and white background. In fact, this setting's prediction accuracy of 97% is very similar to the accuracy (95%, Table 2) of the IAM dataset[22] which were used in benchmarking the TrOCR model to contain other image effects but not color effects. Therefore, it can be seen that the 84.3% accuracy of Case 1 with all image effects randomized and included which is significantly lower than that for IAM (95%) or Case 3 with 97% is due to the existence of other colored fonts and backgrounds which were not in the TrOCR training datasets.

These performance results are limited to the use of the pretrained TrOCR base model. Future work to improve this study could include the fine-tuning of this transformer model with colored text images to enhance the prediction accuracy and employ the TrOCT large model which has 67% more parameters and 1-5% higher accuracy on various benchmark datasets than the base model[21]. In addition, as the primary aim of this work is to explore the individual effects of each image factor with a broader range of contexts that is often more reflective of real-world scenarios where multiple characteristics vary simultaneously, we did not implement sophisticated controls or statistical analysis. But we acknowledge the potential value of formal statistical analyses and plan to incorporate them in future research to investigate the interactions of different factors more rigorously. Moreover, we did not perform more detailed error analysis of different kinds of mistakes such as character confusion and word segmentation errors, as one error was found to be sufficient to reveal the individual image effect based on high correlations of the studied different errors as shown in Table 3 and Fig.4. Nevertheless, it will be interesting to perform a thorough error analysis to provide more details of recognition mistakes in the future.

Overall, the above analysis clearly shows that the font/background color is the single most impactful factor for the TrOCR model's accuracy, which was not studied before. This also suggests that future development of transformer-based OCR models may consider the color effects in training datasets.

Conclusion

This study provides the first systematic, quantitative, and comparative analysis of various image effects of the TrOCR model's performance across diverse datasets, highlighting its strengths and limitations in comprehensive text recognition scenarios. Besides the widely used handwritten dataset IAM, this work has evaluated 10,000 randomized text images. All random effects of font color, background color, font size, font style, blur, noise, and rotation combined cover $>\sim$1.8 million different image effect situations. This research has shed light on how these factors influence the model's performances.

Results of several datasets of all these included image effects completely randomized show similar prediction accuracies, indicating a consistent performance of TrOCR across different random effects and texts. We also found a nearly 100% word detection rate for this most recent OCR model.

Results demonstrate that TrOCR excels in recognizing structured and less varied text scenarios, achieving a high accuracy rate of 99.7% on fixed setting dataset. Its accuracy can be as high as 97% despite all other mentioned image effects being simultaneously included, when the color settings are the same as black/white in its training datasets. Such a strong performance confirms the efficacy of TrOCR models in standard OCR tasks.

Conversely, the accuracy drops significantly when other font and background colors are introduced with these text images. This decline emphasizes the model's sensitivity to color effects which have not been studied before. Blurring effect is the next factor (though much smaller than the color effect) that notably affects the prediction accuracy. In contrast, other studied image effects (font size, font style, noise, and rotation) are not significant in downgrading OCR predictions. Therefore, these results also suggest specific interesting areas for potential improvement of the state-of-the-art TrOCR model to include these significant effects in training datasets to deal with more realistic real-world OCR scenes.

In conclusion, TrOCR represents a significant advancement in OCR technology and can accurately predict texts against various kinds of image effects and texts with $\sim$95% and higher accuracy for black font and white background datasets. Despite its high performance on specific datasets, TrOCR's performance variability across different types of data highlights the need for further fine-tuning and adaptation to handle diverse real-world scenarios effectively to enhance its robustness against varied input conditions. In future work, we plan to fine-tune the TrOCR model on augmented datasets that incorporate a wide range of effects, with a focus on color and blur, to see if this approach can improve the model's robustness and accuracy. By augmenting the training data in this way, we aim to adapt the model to real-world scenarios where such visual distortions are common, potentially enabling it to perform better across various OCR tasks. We will explore the impact of this targeted fine-tuning and report on its effectiveness in addressing the identified limitations in future studies. This research contributes valuable insights into the practical applications of TrOCR and provides a foundation for ongoing developments in the field of handwritten character recognition.

Acknowledgement

The author expresses sincere gratitude to Dr. An Zhao from the University College London for her valuable guidance, assistance, discussion, and feedback on this research. The author also wants to thank various people in Lumiere Education, such as Samatha Silva and Kiana Manian, for their important help in this research and publication process.

References

1 J. Memon, M. Sami and R. Khan, *Handwritten Optical Character Recognition (OCR*, SLR). arXiv:2001.00139v1 [cs.CV.

2 T. Mitchell, *Machine Learning*.

3 Z. Zheng, Y. Zhong, Z. Xiao, W. Lim, S. Tiang, M. Mokayef and C. Wong, *Character Recognition Based on k-Nearest Neighbor, Simple Logistic Regression, and Random Forest*.

4 D. Hijam and S. Saharia, *On developing complete character set meitei mayek handwritten character database*.

5 V. Chauhan, K. Dahiya and A. Sharma, *Problem formulations and solvers in linear SVM: a review*.

6 D. Gowda and V. Kanchana, *Kannada Handwritten Character Recognition and Classification Through OCR Using Hybrid Machine Learning Techniques*.

7 V. Chauhan, S. Singh and A. Sharma, *HCR-Net: A Deep Learning Based Script Independent Handwritten Character Recognition Network*, arXiv:2108.06663 [cs.CV.

8 Y. LeCun, Y. Bengio and G. Hinton, *Deep learning*.

9 P. Verma and G. Foomani, *Improvement in OCR Technologies in Postal Industry Using CNN-RNN Architecture*.

10 C. Boufenar, A. Kerboua and M. Batouche, *Investigation on deep learning for off-line handwritten arabic character recognition*.

11 G. Sokar, E. Hemayed and M. Rehan, *A generic ocr using deep siamese convolution neural networks*.

12 D. Lin, F. Lin, Y. Lv, F. Cai and D. Cao, *Chinese character captcha recognition and performance estimation via deep neural network*.

13 H. Yang, L. Jin and J. Sun, *Recognition of chinese text in historical documents with page-level annotations*.

14 A. Alkaddo and D. Albaqal, *Implementation of OCR using Convolutional Neural Network (CNN): A Surve*.

15 S. Ghasemi and A. Jadidinejad, *Persian text classification via character-level convolutional neural networks*.

16 M. Misgar, F. Mushtaq, S. Khurana and M. Kumar, *Recognition of offline handwritten Urdu characters using RNN and LSTM models*.

17 S. Kulkarni, R. Madurwar, R. Narlawar, A. Pandya, N. Gawande, R. CNN and N.M.F., *Digitization of Physical Notes: A Comprehensive Approach Using OCR*.

18 V. Moulishwar, P. Karthikevan, V. Kumar and A. Srihariharan, *Extension For Handwritten Character Recognition Using Rnn-Gru Algorithm*.

19 D. Diaz, S. Qin, R. Ingle, Y. Fujii and A. Bissacco.

20 P. Ströbel, T. Hodel, W. Boente and M. Volk, *The Adaptability of a Transformer-Based OCR Model for Historical Documents*.

21 M. Li, T. Lv, J. Chen, L. Cui, Y. Lu, D. Florencio, C. Zhang, Z. Li and F. Wei, *TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models*, arXiv: 2109.10282v5 [cs.CL],.

22 U.-V. Marti and H. Bunke, *The IAM-database: an English sentence database for offline handwriting recognition*.

23 R. Atienza, *Vision Transformer for Fast and Efficient Scene Text Recognition*.

24 A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929v2 [cs.CV],.

25 A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, K. Kaiser and I. Polosukhin, *Attention is all you need*.

26 *Training data-efficient image transformers distillation through attention*.

27 H. Bao, L. Dong and F. Wei, BEiT: BERT Pre-Training of Image Transformers. arXiv:2106.08254,.

28 Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, arXiv:1907.11692,.

29 W. Wang, F. Wei, L. Dong, H. Bao, N. Yang and M. Zhou, *Minilm: Deep self-attention distillation for taskagnostic compression of pre-trained transformers*, arXiv:2002.10957,.

30 H. Zhang, E. Whittaker and I. Kitagishi, *Extending TrOCR for Text Localization-Free OCR of Full-Page Scanned Receipt Images*.

31 Y.-H. Chen and P. Strobel, *TrOCR Meets Language Models: An End-to-End Post-correction Approach*, H. Mouchere and A. Zhu (Eds.):.

32 D. Chang and Y. Li.

33 A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala, *PyTorch: An Imperative Style*, arXiv:1912.01703,.

