

Assignment 1- Research Issues 1

Prepared by:
Dr. Nourhan Abdelkerim



Table of content

1. Introduction
2. Title
3. Paper Introduction
4. Problem Statement
5. Importance of the Study
6. Objectives
7. Theoretical Framework & Concepts
8. Literature Review
9. Hypotheses
10. Research Questions
11. Model/Framework Illustration
12. Methodology Summary
13. Data Collection (Sources & Sample)
14. Statistical Methods
15. Contribution & Discussion of Results
16. Conclusion
17. References

1

INTRODUCTION

- This presentation presents a critical analysis of the research paper titled Transformer-Based Optical Character Recognition with Pre-trained Models
- The objective of this assessment is to evaluate the study's adherence to fundamental scientific research standards.
- By systematically examining fourteen key research elements.
- Specifically, we investigate how the authors address the research gap regarding CNN-free backbones in text recognition and how they manipulate independent variables (Transformer architecture) and moderating variables (pre-trained models) to influence the dependent variable (OCR performance).

The title "TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models" is highly relevant to the field of Computer Vision and Natural Language Processing.

- The independent variable is the transformer-based-architecture.
- The dependent variable is the performance in Optical Character Recognition (OCR).
- The moderating variable is Pre-trained models.
- The application field is clearly document digitalization and text recognition.

**TrOCR: Transformer-Based Optical Character Recognition
with Pre-trained Models**

**Minghao Li^{1*}, Tengchao Lv², Jingye Chen^{2*}, Lei Cui²,
Yijuan Lu², Dinei Florencio², Cha Zhang², Zhoujun Li¹, Furu Wei²**

¹Beihang University

²Microsoft Corporation

{liminghao1630, lizj}@buaa.edu.cn

{tengchaolv, v-jingyechen, lecu, yijlu, dinei, chazhang, fuwei}@microsoft.com

The introduction contrasts the traditional method (CNN for image understanding + RNN for text generation) with the proposed method (Transformer for both).

It clearly establishes the variables by defining the problem of text recognition and introducing the shift from hybrid networks to a pure Transformer-based framework.

Abstract

Text recognition is a long-standing research problem for document digitalization. Existing approaches are usually built based on CNN for image understanding and RNN for character-level text generation. In addition, another language model is usually needed to improve the overall accuracy as a post-processing step. In this paper, we propose an end-to-end text recognition approach with pre-trained image Transformer and text Transformer models, namely **TrOCR**, which leverages the Transformer architecture for both image understanding and wordpiece-level text generation. The TrOCR model is simple but effective, and can be pre-trained with large-scale synthetic data and fine-tuned with human-labeled datasets. Experiments show that the TrOCR model outperforms the current state-of-the-art models on the printed, handwritten and scene text recognition tasks. The TrOCR models and code are publicly available at <https://aka.ms/trocr>.

Introduction

Optical Character Recognition (OCR) is the electronic or mechanical conversion of images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo of a document, a scene photo or from subtitle text superimposed on an image. Typically, an OCR system includes two main modules: a text detection module and a text recognition module. Text detection aims to localize all text blocks within the text image, either at word-level or textline-level. The text detection task is usually considered as an object detection problem where conventional object detection models such as YOLOv5 and DBNet (Liao et al. 2019) can be applied. Meanwhile, text recognition aims to understand the text image content and transcribe the visual signals into natural language tokens. The text recognition task is usually framed as an encoder-decoder problem where existing methods leveraged CNN-based encoder for image understanding and RNN-based decoder for text generation. In this paper, we focus on the text recognition task for document images and leave text detection as the future work.

*Work done during internship at Microsoft Research Asia.
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Recent progress in text recognition (Diaz et al. 2021) has witnessed the significant improvements by taking advantage of the Transformer (Vaswani et al. 2017) architectures. However, existing methods are still based on CNNs as the backbone, where the self-attention is built on top of CNN backbones as encoders to understand the text image. For decoders, Connectionist Temporal Classification (CTC) (Graves et al. 2006) is usually used compounded with an external language model on the character-level to improve the overall accuracy. Despite the great success achieved by the hybrid encoder/decoder method, there is still a lot of room to improve with pre-trained CV and NLP models: 1) the network parameters in existing methods are trained from scratch with synthetic/human-labeled datasets, leaving large-scale pre-trained models unexplored. 2) as image Transformers become more and more popular (Dosovitskiy et al. 2021; Touvron et al. 2020), especially the recent self-supervised image pre-training (Bao, Dong, and Wei 2021), it is straightforward to investigate whether pre-trained image Transformers can replace CNN backbones, meanwhile exploiting the pre-trained image Transformers to work together with the pre-trained text Transformers in a single framework on the text recognition task.

To this end, we propose **TrOCR**, an end-to-end Transformer-based OCR model for text recognition with pre-trained CV and NLP models, which is shown in Figure 1. Distinct from the existing text recognition models, TrOCR is a simple but effective model which does not use the CNN as the backbone. Instead, following (Dosovitskiy et al. 2021), it first resizes the input text image into 384×384 and then the image is split into a sequence of 16×16 patches which are used as the input to image Transformers. Standard Transformer architecture with the self-attention mechanism is leveraged on both encoder and decoder parts, where wordpiece units are generated as the recognized text from the input image. To effectively train the TrOCR model, the encoder can be initialized with pre-trained ViT-style models (Dosovitskiy et al. 2021; Touvron et al. 2020; Bao, Dong, and Wei 2021) while the decoder can be initialized with pre-trained BERT-style models (Devlin et al. 2019; Liu et al. 2019; Dong et al. 2019; Wang et al. 2020b), respectively. Therefore, the advantage of TrOCR is three-fold. First, TrOCR uses the pre-trained image Transformer and

The problem is explicitly defined based on a gap: existing methods rely on CNN backbones that introduce image-specific inductive biases and do not fully exploit large-scale pre-trained CV and NLP models.

The "solution" proposed is the TrOCR model, which aims to replace the CNN backbone with an image Transformer.

Recent progress in text recognition (Díaz et al. 2021) has witnessed the significant improvements by taking advantage of the Transformer (Vaswani et al. 2017) architectures. However, existing methods are still based on CNNs as the backbone, where the self-attention is built on top of CNN backbones as encoders to understand the text image. For decoders, Connectionist Temporal Classification (CTC) (Graves et al. 2006) is usually used compounded with an external language model on the character-level to improve the overall accuracy. Despite the great success achieved by the hybrid encoder/decoder method, there is still a lot of room to improve with pre-trained CV and NLP models: 1) the network parameters in existing methods are trained from scratch with synthetic/human-labeled datasets, leaving large-scale pre-trained models unexplored. 2) as image Transformers become more and more popular (Dosovitskiy et al. 2021; Touvron et al. 2020), especially the recent self-supervised image pre-training (Bao, Dong, and Wei 2021), it is straightforward to investigate whether pre-trained image Transformers can replace CNN backbones, meanwhile exploiting the pre-trained image Transformers to work together with the pre-trained text Transformers in a single framework on the text recognition task.

To this end, we propose **TrOCR**, an end-to-end Transformer-based OCR model for text recognition with pre-trained CV and NLP models, which is shown in Figure 1. Distinct from the existing text recognition models, TrOCR is a simple but effective model which does not use the CNN as the backbone. Instead, following (Dosovitskiy et al. 2021), it first resizes the input text image into 384×384 and then the image is split into a sequence of 16×16 patches which are used as the input to image Transformers. Standard Transformer architecture with the self-attention mechanism is leveraged on both encoder and decoder parts, where wordpiece units are generated as the recognized text from the input image. To effectively train the TrOCR model, the encoder can be initialized with pre-trained ViT-style models (Dosovitskiy et al. 2021; Touvron et al. 2020; Bao, Dong, and Wei 2021) while the decoder can be initialized with pre-trained BERT-style models (Devlin et al. 2019; Liu et al. 2019; Dong et al. 2019; Wang et al. 2020b), respectively. Therefore, the advantage of TrOCR is three-fold. First, TrOCR uses the pre-trained image Transformer and

- **Theoretical:** It is the first work to jointly leverage pre-trained image and text Transformers for OCR, removing the need for CNNs and complex heuristics.
- **Applied:** The model is "simple but effective," achieves state-of-the-art results, and is publicly available, making it highly practical for industrial document digitalization.

The contributions of this paper are summarized as follows:

1. We propose TrOCR, an end-to-end Transformer-based OCR model for text recognition with pre-trained CV and NLP models. To the best of our knowledge, this is the first work that jointly leverages pre-trained image and text Transformers for the text recognition task in OCR.
2. TrOCR achieves state-of-the-art results with a standard Transformer-based encoder-decoder model, which is convolution free and does not rely on any complex pre/post-processing steps.
3. The TrOCR models and code are publicly available at <https://aka.ms/trocr>.

TrOCR

Model Architecture

TrOCR is built up with the Transformer architecture, including an image Transformer for extracting the visual features and a text Transformer for language modeling. We adopt the vanilla Transformer encoder-decoder structure in TrOCR. The encoder is designed to obtain the representation of the

The primary objective is to create an end-to-end Transformer-based OCR model that outperforms current SOTA models without complex post-processing. The objectives were realistic and feasible, as demonstrated by the successful experiments on multiple benchmarks (printed, handwritten, and scene text).

text Transformer models, which take advantages of large-scale unlabeled data for image understanding and language modeling, with no need for an external language model. Second, TrOCR does not require any convolutional network for the backbone and does not introduce any image-specific inductive biases, which makes the model very easy to implement and maintain. Finally, experiment results on OCR benchmark datasets show that the TrOCR can achieve state-of-the-art results on printed, handwritten and scene text image datasets without any complex pre/post-processing steps. Furthermore, we can easily extend the TrOCR for multilingual text recognition with minimum efforts, where just leveraging multilingual pre-trained models in the decoder-side and expand the dictionary.

The contributions of this paper are summarized as follows:

1. We propose TrOCR, an end-to-end Transformer-based OCR model for text recognition with pre-trained CV and NLP models. To the best of our knowledge, this is the first work that jointly leverages pre-trained image and text Transformers for the text recognition task in OCR.
2. TrOCR achieves state-of-the-art results with a standard Transformer-based encoder-decoder model, which is convolution free and does not rely on any complex pre/post-processing steps.
3. The TrOCR models and code are publicly available at <https://aka.ms/trocr>.

THEORETICAL FRAMEWORK & CONCEPTS

The paper uses the Transformer architecture (Encoder-Decoder) as its framework. It defines the concepts clearly: the encoder processes image patches (visual features), and the decoder generates word pieces (language modeling). The logic is robust, leveraging the success of Transformers in both vision (ViT) and language (BERT/RoBERTa) tasks to create a unified system.

Recent progress in text recognition (Diaz et al. 2021) has witnessed the significant improvements by taking advantage of the Transformer (Vaswani et al. 2017) architectures. However, existing methods are still based on CNNs as the backbone, where the self-attention is built on top of CNN backbones as encoders to understand the text image. For decoders, Connectionist Temporal Classification (CTC) (Graves et al. 2006) is usually used compounded with an external language model on the character-level to improve the overall accuracy. Despite the great success achieved by the hybrid encoder/decoder method, there is still a lot of room to improve with pre-trained CV and NLP models: 1) the network parameters in existing methods are trained from scratch with synthetic/human-labeled datasets, leaving large-scale pre-trained models unexplored. 2) as image Transformers become more and more popular (Dosovitskiy et al. 2021; Touvron et al. 2020), especially the recent self-supervised image pre-training (Bao, Dong, and Wei 2021), it is straightforward to investigate whether pre-trained image Transformers can replace CNN backbones, meanwhile exploiting the pre-trained image Transformers to work together with the pre-trained text Transformers in a single framework on the text recognition task.

To this end, we propose **TrOCR**, an end-to-end Transformer-based OCR model for text recognition with pre-trained CV and NLP models, which is shown in Figure 1. Distinct from the existing text recognition models, TrOCR is a simple but effective model which does not use the CNN as the backbone. Instead, following (Dosovitskiy et al. 2021), it first resizes the input text image into 384×384 and then the image is split into a sequence of 16×16 patches which are used as the input to image Transformers. Standard Transformer architecture with the self-attention mechanism is leveraged on both encoder and decoder parts, where wordpiece units are generated as the recognized text from the input image. To effectively train the TrOCR model, the encoder can be initialized with pre-trained ViT-style models (Dosovitskiy et al. 2021; Touvron et al. 2020; Bao, Dong, and Wei 2021) while the decoder can be initialized with pre-trained BERT-style models (Devlin et al. 2019; Liu et al. 2019; Dong et al. 2019; Wang et al. 2020b), respectively. Therefore, the advantage of TrOCR is three-fold. First, TrOCR uses the pre-trained image Transformer and

The review is interactive and recent (citing papers from 2017–2021).

It discusses the dominance of CRNN/CTC methods and notes their limitations (complexity, lack of pre-training) to justify the new approach.

It acknowledges where previous methods succeeded but highlights the specific disagreement (need for CNNs) that the paper challenges.

Recent progress in text recognition (Diaz et al. 2021) has witnessed the significant improvements by taking advantage of the Transformer (Vaswani et al. 2017) architectures. However, existing methods are still based on CNNs as the backbone, where the self-attention is built on top of CNN backbones as encoders to understand the text image. For decoders, Connectionist Temporal Classification (CTC) (Graves et al. 2006) is usually used compounded with an external language model on the character-level to improve the overall accuracy. Despite the great success achieved by the hybrid encoder/decoder method, there is still a lot of room to improve with pre-trained CV and NLP models: 1) the network parameters in existing methods are trained from scratch with synthetic/human-labeled datasets, leaving large-scale pre-trained models unexplored. 2) as image Transformers become more and more popular (Dosovitskiy et al. 2021; Touvron et al. 2020), especially the recent self-supervised image pre-training (Bao, Dong, and Wei 2021), it is straightforward to investigate whether pre-trained image Transformers can replace CNN backbones, meanwhile exploiting the pre-trained image Transformers to work together with the pre-trained text Transformers in a single framework on the text recognition task.

The implicit main hypothesis is that a pure Transformer architecture pre-trained on large-scale data can outperform hybrid CNN-RNN models in text recognition.

This links directly to the title and variables by testing the "Transformer-based" independent variable against the standard benchmarks.

Recent progress in text recognition (Diaz et al. 2021) has witnessed the significant improvements by taking advantage of the Transformer (Vaswani et al. 2017) architectures. However, existing methods are still based on CNNs as the backbone, where the self-attention is built on top of CNN backbones as encoders to understand the text image. For decoders, Connectionist Temporal Classification (CTC) (Graves et al. 2006) is usually used compounded with an external language model on the character-level to improve the overall accuracy. Despite the great success achieved by the hybrid encoder/decoder method, there is still a lot of room to improve with pre-trained CV and NLP models: 1) the network parameters in existing methods are trained from scratch with synthetic/human-labeled datasets, leaving large-scale pre-trained models unexplored. 2) as image Transformers become more and more popular (Dosovitskiy et al. 2021; Touvron et al. 2020), especially the recent self-supervised image pre-training (Bao, Dong, and Wei 2021), it is straightforward to investigate whether pre-trained image Transformers can replace CNN backbones, meanwhile exploiting the pre-trained image Transformers to work together with the pre-trained text Transformers in a single framework on the text recognition task.

The core research question addressed is: Can pre-trained image Transformers replace CNN backbones in OCR?. Answering this affirmatively directly achieves the objective of establishing TrOCR as a SOTA model.

Recent progress in text recognition (Diaz et al. 2021) has witnessed the significant improvements by taking advantage of the Transformer (Vaswani et al. 2017) architectures. However, existing methods are still based on CNNs as the backbone, where the self-attention is built on top of CNN backbones as encoders to understand the text image. For decoders, Connectionist Temporal Classification (CTC) (Graves et al. 2006) is usually used compounded with an external language model on the character-level to improve the overall accuracy. Despite the great success achieved by the hybrid encoder/decoder method, there is still a lot of room to improve with pre-trained CV and NLP models: 1) the network parameters in existing methods are trained from scratch with synthetic/human-labeled datasets, leaving large-scale pre-trained models unexplored. 2) as image Transformers become more and more popular (Dosovitskiy et al. 2021; Touvron et al. 2020), especially the recent self-supervised image pre-training (Bao, Dong, and Wei 2021), it is straightforward to investigate whether pre-trained image Transformers can replace CNN backbones, meanwhile exploiting the pre-trained image Transformers to work together with the pre-trained text Transformers in a single framework on the text recognition task.

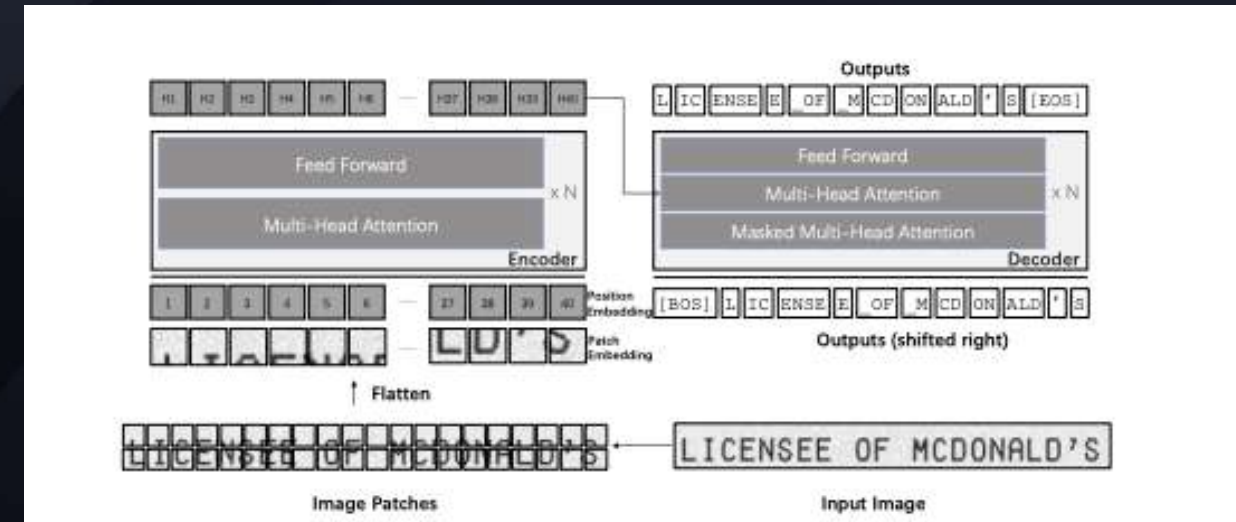
Figure 1 clearly illustrates the model framework.

It shows the flow:

Input Image → Image Patches →

Transformer Encoder → Transformer Decoder → Output Text.

This visualizes the relationship between the visual and textual variables.



The methodology is quantitative and experimental.

It involves a two-stage pre-training process (large-scale synthetic data) followed by fine-tuning on specific task datasets.

It uses an encoder-decoder network structure.

Task Pipeline

In this work, the pipeline of the text recognition task is that given the textline images, the model extracts the visual features and predicts the wordpiece tokens relying on the image and the context generated before. The sequence of ground truth tokens is followed by an "[EOS]" token, which indicates the end of a sentence. During training, we shift the sequence backward by one place and add the "[BOS]" token to the beginning indicating the start of generation. The shifted ground truth sequence is fed into the decoder, and the output of that is supervised by the original ground truth sequence with the cross-entropy loss. For inference, the decoder starts from the "[BOS]" token to predict the output iteratively while continuously taking the newly generated output as the next input.

Pre-training

We use the text recognition task for the pre-training phase, since this task can make the models learn the knowledge of both the visual feature extraction and the language model. The pre-training process is divided into two stages that differ by the used dataset. In the first stage, we synthesize a large-scale dataset consisting of hundreds of millions of printed textline images and pre-train the TrOCR models on that. In the second stage, we build two relatively small datasets corresponding to printed and handwritten downstream tasks, containing millions of textline images each. We use the existed and widely adopted synthetic scene text datasets for the scene text recognition task. Subsequently, we pre-train separate models on these task-specific datasets in the second stage, all initialized by the first-stage model.

Fine-tuning

Except for the experiments regarding scene text recognition, the pre-trained TrOCR models are fine-tuned on the downstream text recognition tasks. The outputs of the TrOCR models are based on Byte Pair Encoding (BPE) (Sennrich, Haddow, and Birch 2015) and SentencePiece (Kudo and Richardson 2018) and do not rely on any task-related vocabularies.

Data Augmentation

We leverage data augmentation to enhance the variety of the pre-training and fine-tuning data. Six kinds of image trans-

- Sources: Synthetic data generated via TRDG, public PDF files, and standard datasets (SROIE, IAM, IIIT5K, etc.).
- Sample: First-stage pre-training used 684 million text lines from PDFs. Second-stage used millions of handwritten and printed lines.
- Method: Crawling PDFs, synthesizing text, and using established benchmark datasets.

Experiments

Data

Pre-training Dataset To build a large-scale high-quality dataset, we sample two million document pages from the publicly available PDF files on the Internet. Since the PDF files are digital-born, we can get pretty printed textline images by converting them into page images and extracting the textlines with their cropped images. In total, the first-stage pre-training dataset contains 684M textlines.

We use 5,427 handwritten fonts¹ to synthesize handwritten textline images by the TRDG², an open-source text recognition data generator. The text used for generation is crawled from random pages of Wikipedia. The handwritten dataset for the second-stage pre-training consists of 17.9M textlines, including IIIT-HWS dataset (Krishnan and Jawahar 2016). In addition, we collect around 53K receipt images in the real world and recognize the text on them by commercial OCR engines. According to the results, we crop the textlines by their coordinates and rectify them into normalized images. We also use TRDG to synthesize 1M printed textline images with two receipt fonts and the built-in printed fonts. In total, the printed dataset consists of 3.3M

The paper evaluates performance using standard metrics: Precision, Recall, F1 Score (for SROIE), Character Error Rate (CER) (for IAM), and Word Accuracy (for scene text).

It compares these statistics against baseline models like CRNN and Tesseract.

Evaluation Metrics

The SROIE dataset is evaluated using the word-level precision, recall and f1 score. If repeated words appear in the ground truth, they are also supposed to appear in the prediction. The precision, recall and f1 score are described as:

$$Precision = \frac{\text{Correct matches}}{\text{The number of the detected words}}$$

$$Recall = \frac{\text{Correct matches}}{\text{The number of the ground truth words}}$$

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The IAM dataset is evaluated by the case-sensitive Character Error Rate (CER). The scene text datasets are evaluated by the Word Accuracy. For fair comparison, we filter the final output string to suit the popular 36-character charset (lowercase alphanumeric) in this task.

CONTRIBUTION & DISCUSSION OF RESULTS

- Results: TrOCR outperformed SOTA models on SROIE and IAM datasets, and achieved new SOTA on 5 out of 8 scene text datasets.
- Discussion: The results confirm that Transformer decoders are more competitive than CTC decoders and do not require external language models.
- Achievement: The objectives were fully achieved.
- Recommendations: The paper suggests the model is suitable for industrial use due to speed/accuracy and proposes future work in text detection.

Results

Architecture Comparison We compare different combinations of the encoder and decoder to find the best settings. For encoders, we compare DeiT, BEiT and the ResNet-50 network. Both the DeiT and BEiT are the base models in their original papers. For decoders, we compare the base decoders initialized by RoBERTa_{BASE} and the large decoders initialized by RoBERTa_{LARGE}. For further comparison, we also evaluate the CRNN baseline model and the Tesseract OCR in this section, while the latter is an open-source OCR Engine using the LSTM network.

IAM Handwriting Database Table 4 shows the results of the TrOCR models and the existing methods on the IAM Handwriting Database. According to the results, the methods with CTC decoders show good performance on this task and the external LM will result in a significant reduction in CER. By comparing the methods (Bluche and Messina 2017) with the TrOCR models, the TrOCR_{LARGE} achieves a better result, which indicates that the Transformer decoder is more competitive than the CTC decoder in text recognition and has enough ability for language modeling instead of relying on an external LM. Most of the methods use sequence models in their encoders after the CNN-based backbone except the FCN encoders in (Wang et al. 2020a), which leads to a significant improvement on CER. Instead of relying on the features from the CNN-based backbone, the TrOCR models using the information from the image patches get similar and even better results, illustrating that the Transformer structures are competent to extract visual features well after

Inference Speed Table 5 shows the inference speed of different settings TrOCR models on the IAM Handwriting Database. We can conclude that there is no significant margin in inference speed between the base models and the large models. In contrast, the small model shows comparable results for printed and handwriting text recognition even though the number of parameters is an order of magnitude smaller and the inference speed is as twice as fast. The low number of parameters and high inference speed means fewer computational resources and user waiting time, making it more suitable for deployment in industrial applications.

16. Conclusion

The application of the Research Issues course criteria demonstrates that the TrOCR paper represents a highly structured and methodological piece of experimental research.

The study successfully identifies a clear research gap the reliance on complex CNN-RNN hybrids and formulates a solution that is logically consistent with its title and objectives.

The analysis highlights several strengths in the paper's design:

1. **Variable Clarity:** The relationship between the independent variable (Transformer architecture), the moderating variable (pre-trained models), and the dependent variable (accuracy/speed) is explicitly defined and visualized in the framework.
2. **Methodological Rigor:** The use of large-scale synthetic data collection combined with standard statistical metrics (CER, F1 Score) ensures the reliability of the results.
3. **Contribution:** The study fulfills its theoretical importance by proving that pure Transformers can outperform hybrid models, and its applied importance by offering a model suitable for industrial deployment due to its efficiency.

Ultimately, the paper fulfills all fourteen assessment points effectively, validating its hypotheses and offering clear recommendations for future research in document digitalization.

17. References

- <https://ojs.aaai.org/index.php/AAAI/article/view/26538>



Thank You