Conference Paper Title*

*Note: Sub-titles are not captured in Xplore and should not be used

1st Yilin Chen
Department of Statistics
University of Michigan Ann Arbor
Ann Arbor, United Stated of America
yilincyl@umich.edu

***Abstract*—This project explores the application of a transformer-based OCR model, TrOCR, to the task of automatic license plate recognition (ALPR). By fine-tuning Microsoft's pre-trained trocr-base-handwritten model on the UK-Car-Plate-VRN dataset, the project evaluates the feasibility of leveraging sequence-to-sequence transformer architectures for recognizing real-world license plates in varying visual conditions. Results show promising character-level accuracy, although numerical transcriptions remain a challenge due to limited digit exposure in training. The study demonstrates that OCR models can be adapted with minimal labeled data and lays the groundwork for future real-time ALPR systems.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

Automatic License Plate Recognition (ALPR) is a computer vision task that plays a pivotal role in intelligent transportation systems, surveillance, and access control. Traditionally, ALPR systems consist of multiple components, including object detection, segmentation, character recognition, and rule-based post-processing. However, recent advancements in deep learning—particularly Transformer-based models—enable end-to-end architectures that significantly reduce complexity.

TrOCR (Transformer-based Optical Character Recognition) by Microsoft leverages a Vision Transformer (ViT) encoder and a language modeling decoder to perform end-to-end sequence transduction. Fine-tuning such pre-trained models has shown success in a variety of document OCR tasks. This project investigates whether TrOCR can be repurposed for license plate transcription using a small-scale, real-world dataset.

II. Method

We approach the ALPR problem as a supervised image-to-sequence task. The model input consists of a cropped license plate image, and the output is the alphanumeric plate number as a string.

We use microsoft/trocr-base-handwritten, a pre-trained VisionEncoderDecoder model. The encoder is a Vision Transformer that encodes the image into visual features, while the decoder is a causal language model that predicts text tokens autoregressively. We modify the label format to align with TrOCR expectations: text is converted to JSON-like strings (e.g., "plate": "BD75YGT").

The dataset used is the Hugging Face spawn99/UK-Car-Plate-VRN-Dataset, which includes front license plate images and associated Vehicle Registration Numbers (VRNs). To accommodate hardware limitations, we trained on a subset of 50–100 images with small batch sizes (1–4) and a learning rate tuned with early stopping. All images were resized and normalized before being fed into the model.

III. Results

The model achieved rapid convergence on a small training set, with training loss decreasing to below 2.0 in fewer than three epochs. Evaluation on unseen samples revealed reasonable character-level accuracy but exposed systematic substitution errors—particularly digits being misread as alphabetic characters (e.g., "75" → "YY"). These failures can be attributed to the model's limited exposure to numerical patterns.

To address this, we tried to implemented beam search decoding with regular expression filtering to extract strings matching UK license formats (e.g., [A-Z]2[0-9]2[A-Z0-9]3). The decoded results improved in structural validity but still suffered from semantic ambiguities. And this further decrease the accuracy of recognizing plates in other form.

A. Qualitative predictions:

Input: [Image of ABCDE] → Predicted: AADEDEE
Input: [Image of BD75YGT] → Predicted: BDYYYYYYYYYY

These results show that while TrOCR can model plate text structures, but due to the limitation of the training time, the model was unable to train based on a large dataset. Also the TrOCR have limited recognition on numbers of a plate.

IV. Conclusion

This work demonstrates that TrOCR, though originally trained for handwritten documents, can be adapted to real-world ALPR tasks with minimal supervision. Despite limitations in number recognition and generalization due to dataset size, the model performs well in learning character sequence structure. Future work includes training on a larger set with

Identify applicable funding agency here. If none, delete this.

balanced alphanumeric distribution, synthetic image augmentation for digits, and potential integration with YOLO-based plate detectors to enable full ALPR pipelines.

Additionally, incorporating character-specific loss metrics and weighted sampling could help alleviate the character imbalance issue. TrOCR's flexibility in accepting task-specific prompts makes it a strong candidate for further adaptation across OCR domains.

REFERENCES

[1] i, M., et al. (2021). TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models. arXiv preprint arXiv:2109.10282

[2] Dosovitskiy, A., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

[3] He, K., et al. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE CVPR.

[4] Hugging Face. spawn99/UK-Car-Plate-VRN-Dataset. https://huggingface.co/datasets/spawn99/UK-Car-Plate-VRN-Dataset