TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models

Minghao Li[1]*, **Tengchao Lv**[2], **Lei Cui**[2], **Yijuan Lu**[3],
Dinei Florencio[3], **Cha Zhang**[3], **Zhoujun Li**[1], **Furu Wei**[2]
[1]Beihang University
[2]Microsoft Research Asia
[3]Microsoft Azure AI
{liminghao1630,lizj}@buaa.edu.cn
{tengchaolv,lecu,yijlu,dinei,chazhang,fuwei}@microsoft.com

Abstract

Text recognition is a long-standing research problem for document digitalization. Existing approaches for text recognition are usually built based on CNN for image understanding and RNN for char-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 both printed and handwritten text recognition

content and transcribe the visual signals into natural language tokens. The text recognition task is usually framed as an encoder-decoder problem where existing approaches 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.

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 approaches 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