Abstract

This work proposes a fully Transformers-based method in the area of Scene Text Recognition. The key idea of the work lies in the use of multi-modal cues. We explore a two-staged training strategy to train our model. During the pretraining stage, a masking strategy is applied to help learn multi-modal features, and a semi-supervised method is proposed to enable introduce unlabeled real data. In the fine-tuning stage, we use an iterative correction method to improve the performance. As shown in the figure, our model, MVLT, successfully recognizes texts in complex real-world scenarios.

Methods

MVLT is built within an encoder-decoder architecture, with a Vision Transformer (ViT) encoder, and a multi-modal Transformer decoder. In pretraining, a part of image patches are masked and taken by the encoder as the input. In the meantime, one sub-decoder takes image features and partially masked character embeddings as input, while another sub-decoder takes image features and totally masked character embeddings as input. The purpose of pretraining is to rebuild the masked image patches and predict the masked characters, by which endowing the model with the ability to recognize text using multi-modal cues.

In the fine-tuning stage, all of the image patches are visible. Before the first iterative correction starts, the decoder takes the image feature and a sequence of mask token embeddings as input. When the iterative correction starts, the predicted character embeddings of the current iteration are taken as the input by the decoder in the next iteration.

Results

We compare the proposed method with several strong baselines, including ASTER[2], SRN[3], and ABINet[4] on 6 commonly used test datasets. MVLT reaches SOTA performance, and MVLT* is superior.

References: