Masked Vision-Language Transformers for Scene Text Recognition

Jie Wu*, Ying Peng, Shengming Zhang, Weigang Qi, Jian Zhang

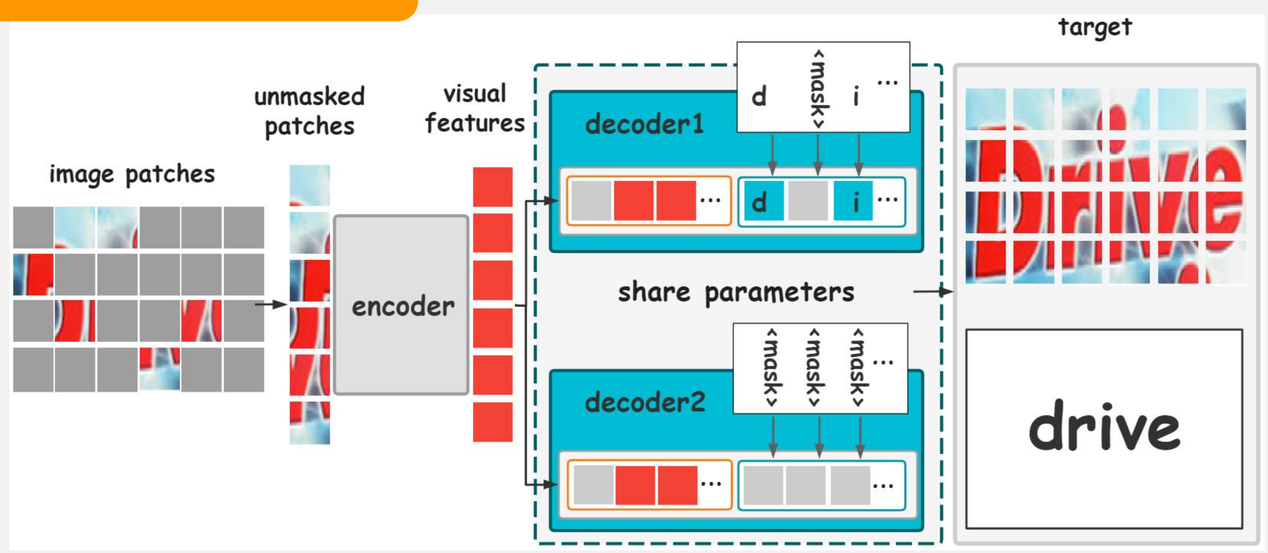
Westone Information Industry INC. Chengdu, China

* corresponding author: wu.jie@westone.com.cn

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) [1]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.

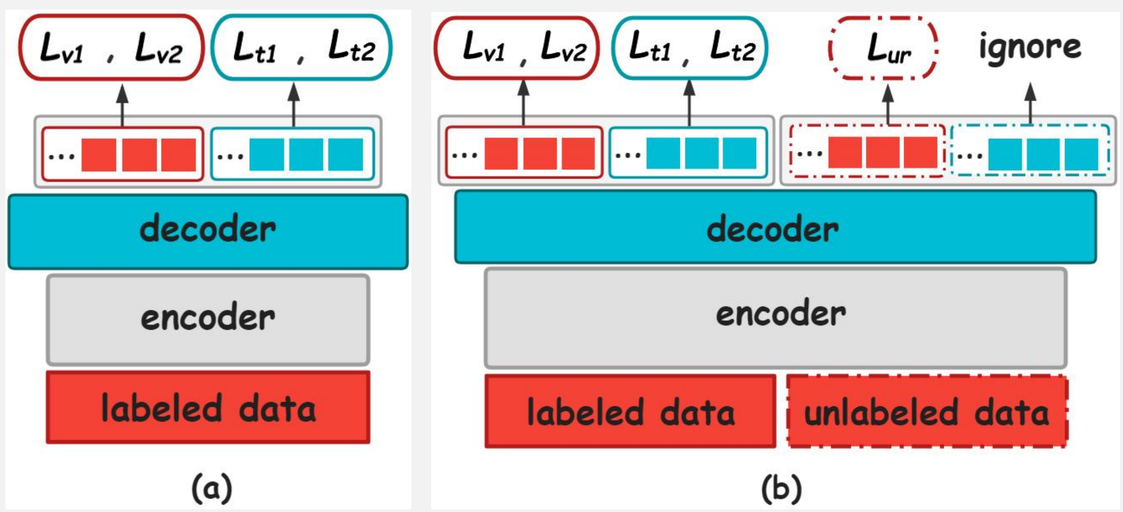
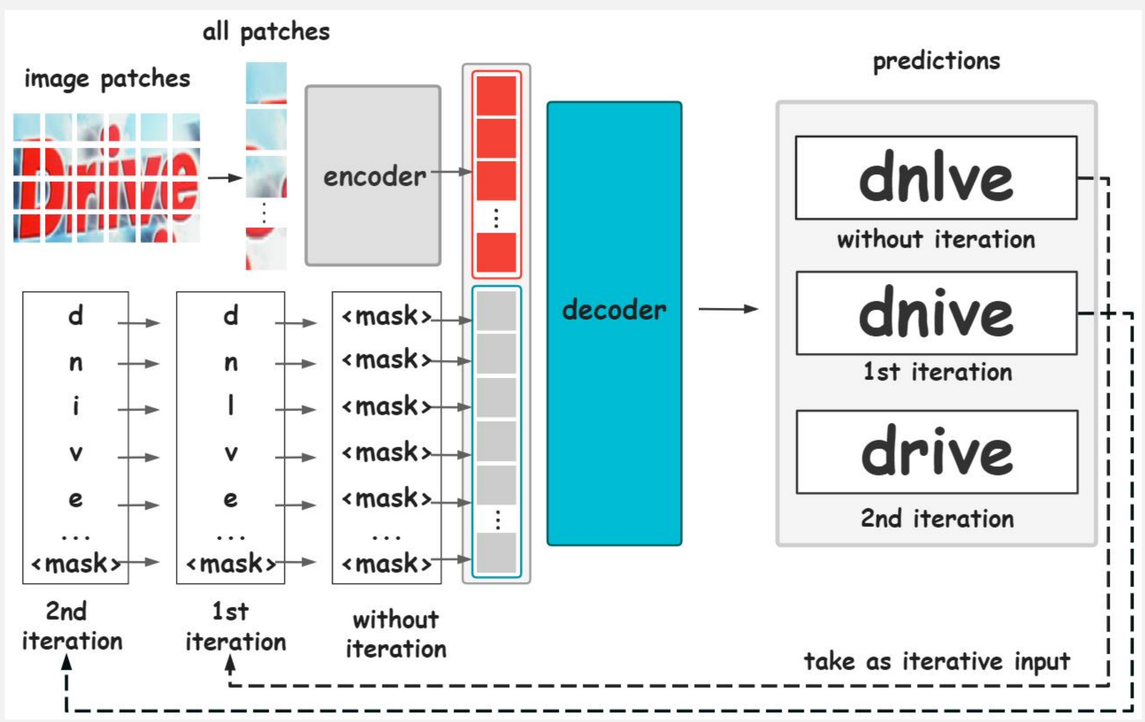
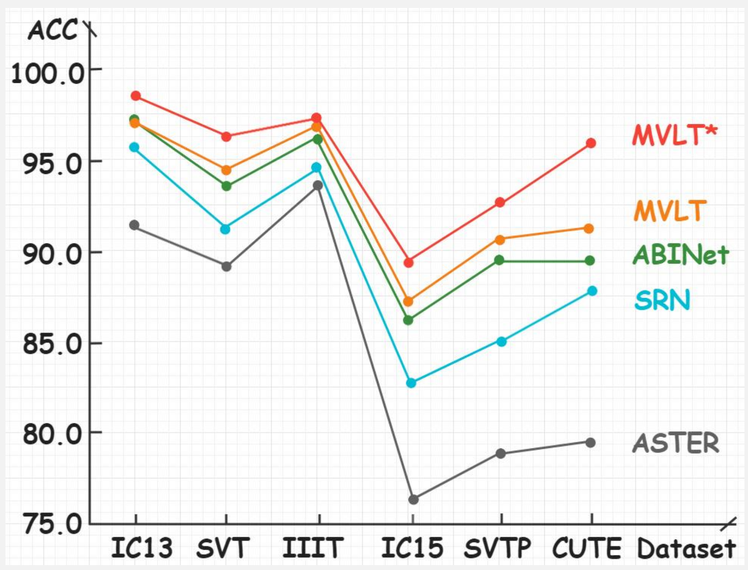


Figure (a) illustrates training MVLT by using labeled data, and Figure (b) illustrates MVLT*, which uses additional **unlabeled real** data in pretraining. L_{v1}, L_{v2}, and L_{ur} are losses related to re-build image patches, and L_{t1} and L_{t2} related to predict characters. We concatenate labeled and unlabeled data along the batch dimension as the input, and when computing the loss corresponding to the unlabeled data, we ignore text-related losses.



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.

Reference:

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Syl_x0002_vain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Trans_x0002_formers for image recognition at scale. In 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[2] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. ASTER: an attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell., 41(9):2035–2048, 2019.

[3] Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020, pages 12110–12119. Computer Vision Foundation / IEEE, 2020.

[4] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In IEEE Conference on Computer Vision and Pattern Recog_x0002_nition, CVPR 2021, virtual, June 19-25, 2021, pages 7098–7107. Computer Vision Foundation / IEEE, 2021.