Scene Text Recognition via Transformer

Authors: Xinjie Feng, Hongxun Yao, Yuankai Qi, Jun Zhang, Shengping Zhang

October 20, 2020

Presented By: Oladotun Aluko

Outline

- Introduction
- Approaches to Scene Text Recognition
- Proposed Method
- Architecture
- Experiments and Results
- References

Introduction

- Scene text is the text that appears in an image captured by a camera in an outdoor environment. The text in scene images varies in shape, font, color, and position
- Scene text recognition can be roughly grouped into two categorizations:
  - text recognition with regular shapes and
  - text recognition with arbitrary shapes

Approaches to Scene Text Recognition

Effective methods to approach scene text recognition with arbitrary shapes can be roughly grouped into three categories:

- rectification based methods,
- segmentation based methods and
- spatial attention-based methods

Challenge with Scene Text Recognition

(a) slight bend or lean

(b) moderate bend or lean

(c) severe bend or lean

Proposed Method

Category of spatial attention-based method. The proposed model consists of two major modules:

- a feature extractor module and
- transformer module

Architecture

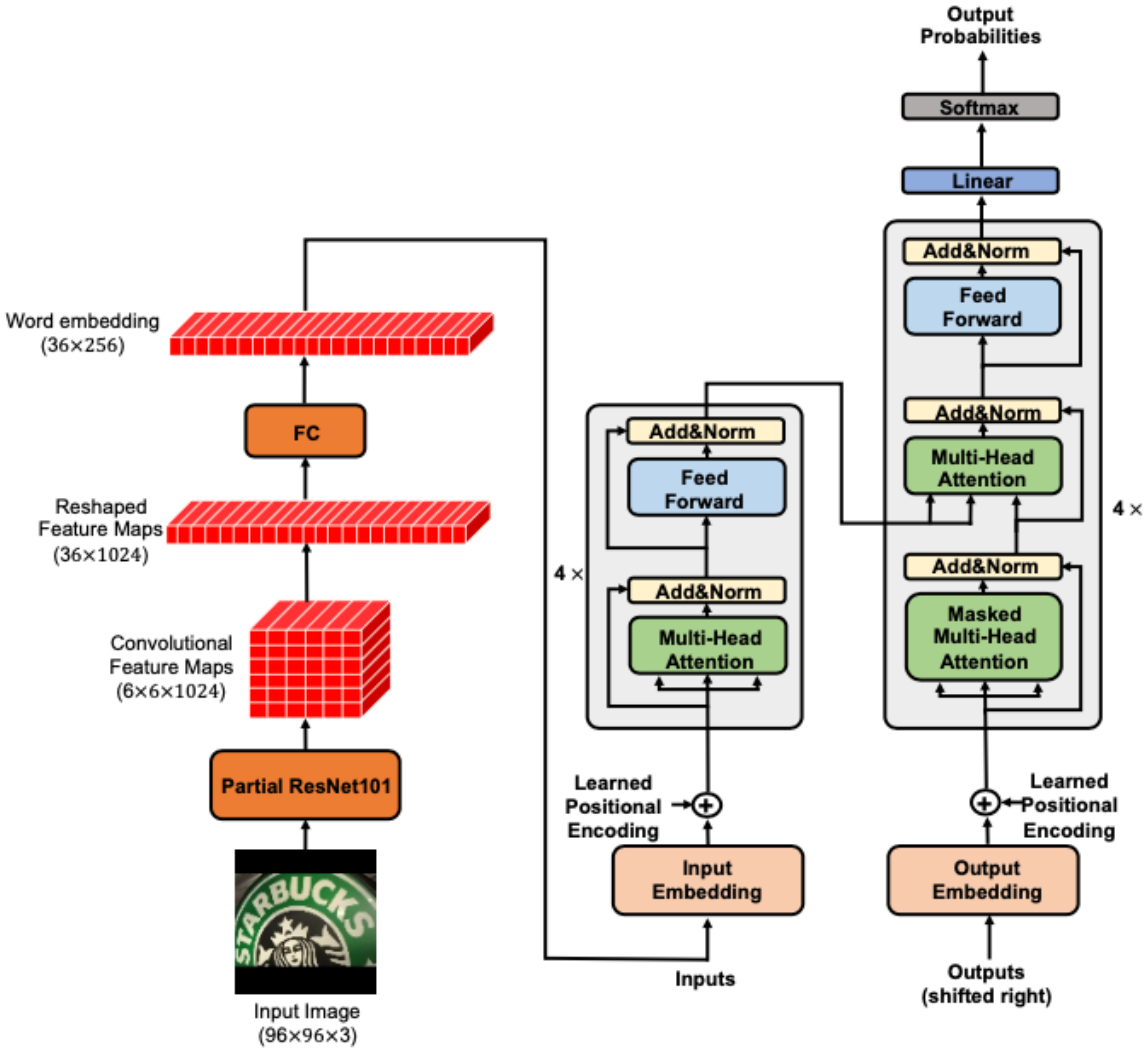


Fig. 2. Overview of the proposed Transformer OCR, which contains two cascade modules: the feature extractor module and the transformer module.

Architecture(contd)

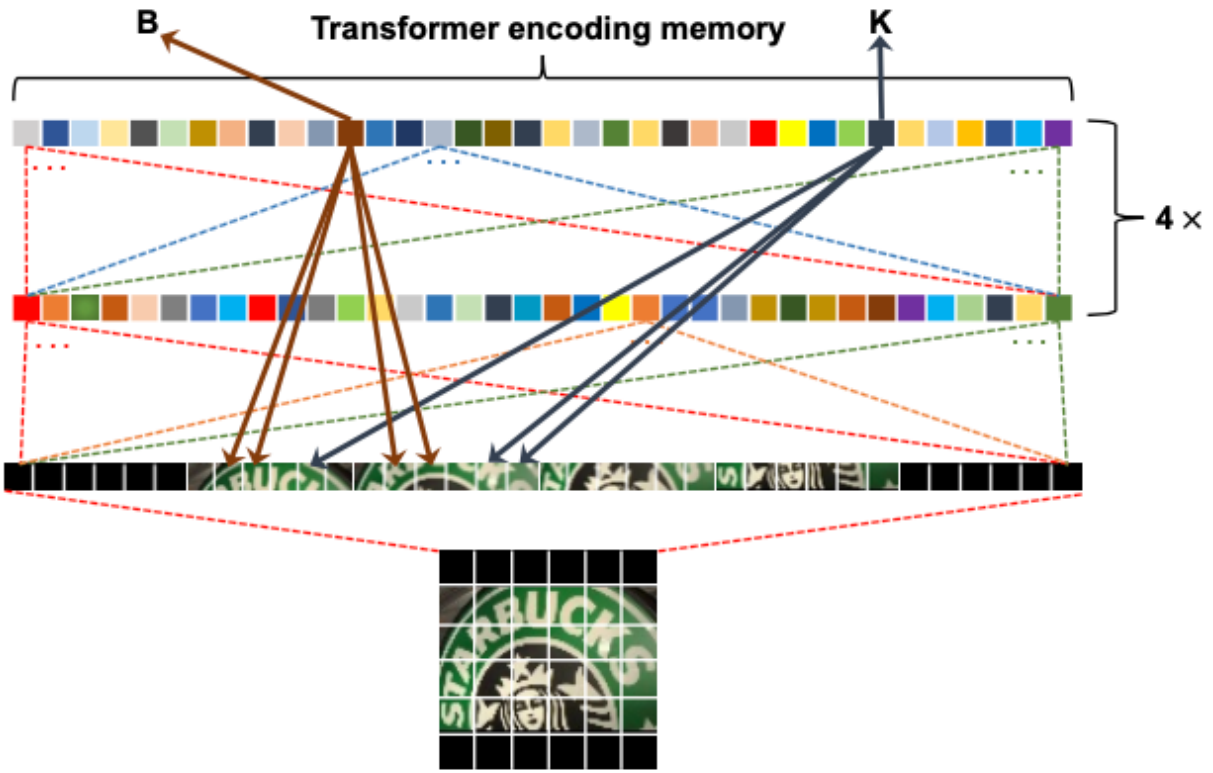


Fig. 3. Spatial attention realized by cross attention mechanism of transformer encoder.

Experiments and Results

Evaluations are based on 5 datasets:

- IIIT5K-Words (IIIT5K)
- Street View Text (SVT)
- ICDAR 2015 (IC15)
- SVT-Perspective (SVTP)
- CUTE80 (CUTE)

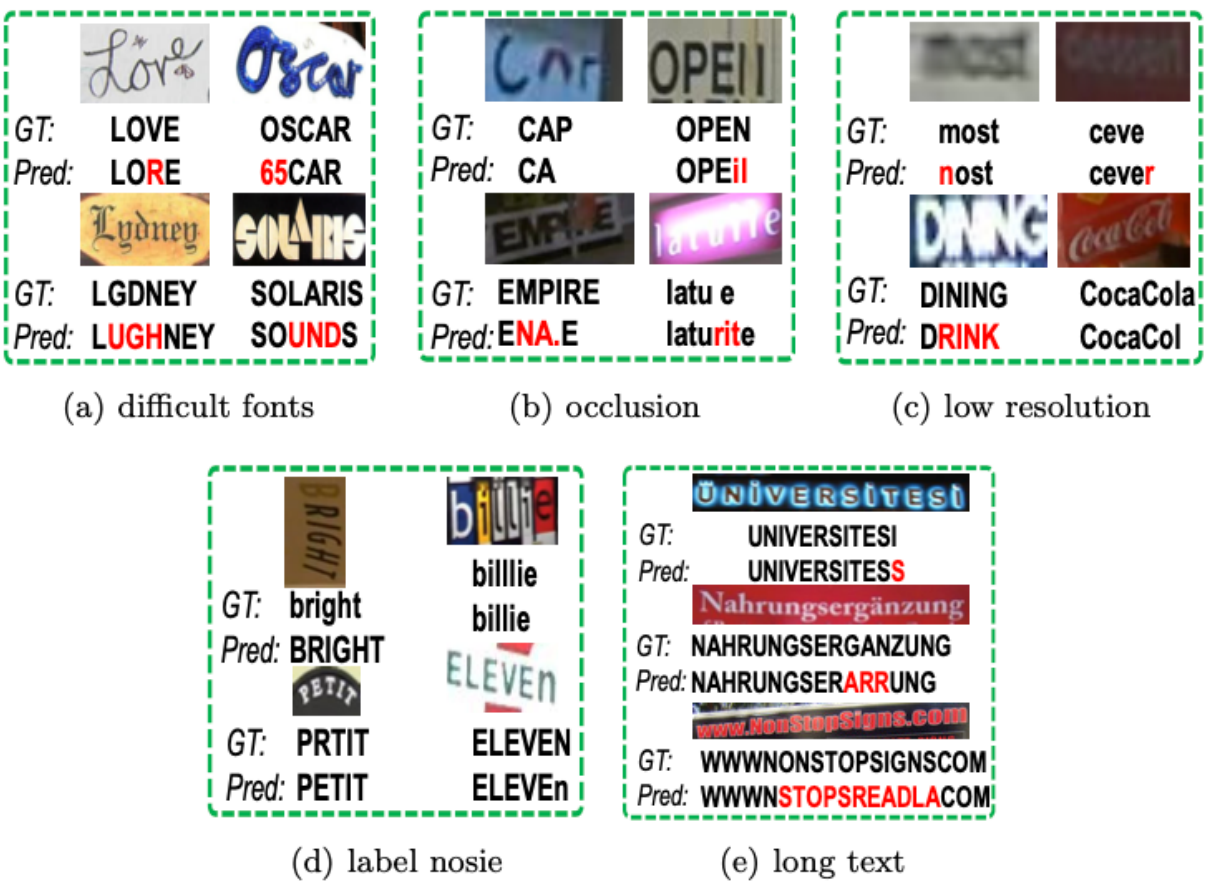
Experiments and Results(contd)

Methods	IIIT5k			SVT		IC15	SVTP	CUTE
	50	1k	0	50	0	0	0	0
Wang et al. [51]	-	-	-	57.0	-	-	-	-
Mishra et al. [36]	64.1	57.5	-	73.2	-	-	-	-
Wang et al. [53]	-	-	-	70.0	-	-	-	-
Almazan et al. [1]	91.2	82.1	-	89.2	-	-	-	-
Yao et al. [60]	80.2	69.3	-	75.9	-	-	-	-
Rodríguez et al. [40]	76.1	57.4	-	70.0	-	-	-	-
Jaderberg et al. [23]	-	-	-	86.1	-	-	-	-
Su and Lu [47]	-	-	-	83.0	-	-	-	-
Gordo [14]	93.3	86.6	-	91.8	-	-	-	-
Jaderberg et al. [21]	97.1	92.7	-	95.4	80.7	-	-	-
Jaderberg et al. [20]	95.5	89.6	-	93.2	71.7	-	-	-
Shi et al. [42]	97.8	95.0	81.2	97.5	82.7	-	-	-
Shi et al. [43]	96.2	93.8	81.9	95.5	81.9	-	71.8	59.2
Lee et al. [27]	96.8	94.4	78.4	96.3	80.7	-	-	-
Yang et al. [58]	97.8	96.1	-	95.2	-	-	75.8	69.3
Cheng et al. [7]	99.3	97.5	87.4	97.1	85.9	70.6	-	-
Cheng et al. [8]	99.6	98.1	87.0	96.0	82.8	68.2	73.0	76.8
Liu et al. [30]	-	-	92.0	-	85.5	74.2	78.9	-
Bai et al. [4]	99.5	97.9	88.3	96.6	87.5	73.9	-	-
Liu et al. [31]	97.0	94.1	87.0	95.2	-	-	-	-
Liu et al. [32]	97.3	96.1	89.4	96.8	87.1	-	73.9	62.5
Yang et al. [59]	97.8	96.1	-	95.2	-	-	75.8	69.3
Liao et al. [29]	99.8	98.8	91.9	98.8	86.4	-	-	79.9
Shi et al. [44]	99.6	98.8	93.4	97.4	89.5	76.1	78.5	79.5
Yang et al [56]	-	-	94.2	-	89.0	74.8	81.7	83.7
Yang et al. [57]	99.5	98.8	94.4	97.2	88.9	78.7	80.8	87.5
Lyu et al. [34]	99.8	99.1	94.0	97.2	90.1	76.3	82.3	86.8
Liao et al. [33]	99.8	99.3	95.3	99.1	91.8	78.2	83.6	88.5
Li et al. [28]	99.4	98.2	95.0	98.5	91.2	78.8	86.4	89.6
Our method	**99.8**	**99.5**	**98.1**	**99.1**	**98.6**	**90.3**	**98.2**	**99.3**

Method Failure Cases

- The method fails in some challenging cases such as difficult fonts, occlusion, or low resolution
- The method also struggles to recognize the long texts

Method Failure Cases (contd)



(a) difficult fonts (b) occlusion (c) low resolution

(d) label nosie (e) long text

Fig. 5. Some failed cases of our method.

References

1. Scene text recognition via Transformer
2. Colab Notebook