International Core Journal of Engineering Volume 8 Issue 2, 2022
ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0059

Text Recognition Method based on Transformer

Bin Cao, Zhijiang Bai

Shanghai Maritime University, Shanghai 201306, China

Abstract

Text with advanced semantics can be seen everywhere in our life, and character recognition is an important research direction in the field of computer vision.However, in real life, text is faced with complex text background and various types of text. This paper focuses on the text images of natural scenes and proposes a method to recognize the text of natural scenes based on Transformer model.This method achieves good results on ICDAR2015 data set.

Keywords

Text Recogntion; Transformer; Deep Learning; Image.

1. Introduction

As the carrier of cultural inheritance, text are the symbol of human civilization.It is also an important tool for human beings to understand the world.The high-level semantics generated by complementary text and image background are of great significance to the process of human cognition of the world.Images usually contain a lot of text information, which plays an important role in understanding the content of images.Nowadays, with the rapid development of artificial intelligence and big data technology, more and more fields need to use text information in images.Character recognition is an advanced topic in current research.In practical application, the text on the image has complex problems such as fuzzy font and low definition.

Jaderberg [1]Et al. proposed an end-to-end text recognition method combining text detection based on regional suggestion mechanism and text recognition based on CNN, which made a great breakthrough and contribution to text recognition.Liu, etc.[2] A FOTS model for end-to-end word detection and recognition is proposed. The proposed method introduces the rotation sensing region to share the convolution feature between detection and recognition, which has a higher speed.Luo etc.[3]An end-to-end scene text recognition network is proposed, which is composed of MORN, a rectification sub-network, and ASRN, a recognition subnet. MORN and ASRN can carry out end-to-end joint learning, and the training process does not require the supervision information of mark character position or pixel-level segmentation, but the recognition effect of long words is not good.Qin etc.[4] An end - to - end scene text recognition network for arbitrary shape text recognition is proposed.In this method, arbitrary shape word detection is reduced to instance segmentation, and then the arbitrary shape word region is decoded by using attentional mechanism model.Subsequently, Sarshogh M Rd et al[5] This paper proposes a multi - task end - to - end recognition network which can simultaneously detect, recognize and classify characters.

2. Method

The proposed model consists of two parts: feature extraction network and Transformer module.The Transformer structure is shown in Figure 1:

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0059

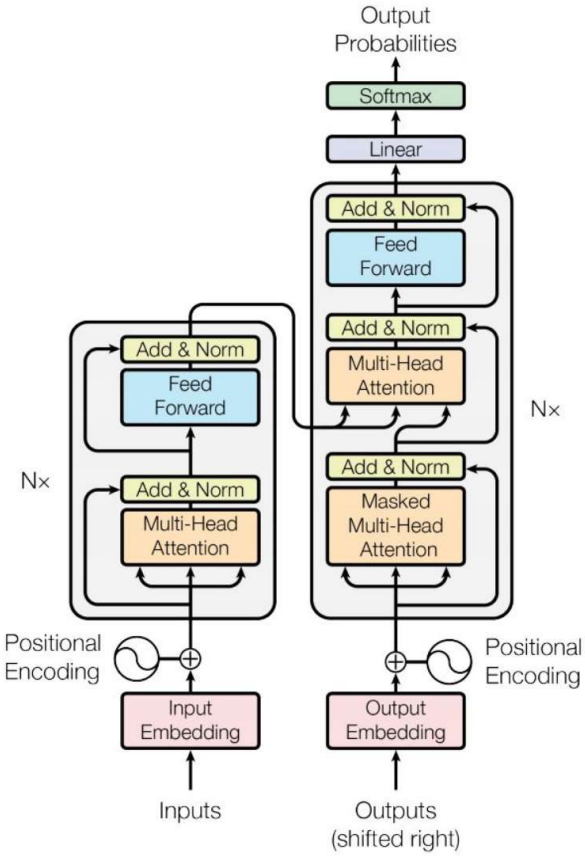

Figure 1. The structure of the transformer

Transformer model is roughly divided into Encoder and Decoder, corresponding to the left and right parts in the figure above respectively.The encoder is stacked with N identical layers (N=6 in our later experiments), and each layer has two sub-layers.The first sublayer is a multi-head Attention, and the second sublayer is a simple Feed Forward.A residual join + Layer normalization operation has been added to both sub-layers.The model's decoder also stacks N identical layers, but each layer has a slightly different structure than in the encoder.For each layer of the decoder, besides the two sub-layers of multi-head Attention and Feed Forward in the encoder, the decoder also contains a sub-layer of Masked multi-head Attention, As shown in the figure, there is also residual and Layer normalization for each of the sub-layers.The Input of the model is composed of two parts: Input Embedding and Positional Encoding, and the output of the model is simply obtained by Softmax through the output of Decoder.In this model, pretrained resnet-18 is used to extract image features. Backbone will output a feature map with dimension [BATch_size, 512, 1, 24] when batch_size is ignored. Each image will get a 1×24 feature graph with 512 channels as shown below. The feature values at the same position of different channels are splicted to form a new vector, which is used as a time step input. At this time, the input with dimension [BATch_size, 24, 512] is constructed. In the end, greedy algorithm decoding is used to predict OCR results directly.Because the model only produces one output at a time, the character corresponding to the highest probability in the probability distribution of the output is selected as the result of this prediction, and then the next character is predicted.

ResNet (Residual Neural Network) was used for feature extraction. ResNet (Residual Neural Network) was proposed by He et al.[6], successfully trained 152-layer Neural Network by using Residual block structure, and won the first prize in the classification task of ImageNet competition in 2015.The error rate of top5 is 3.57%, and the number of references is low.The residual block adds an identity map by means of skip connection structure, and combines the feature information of the previous layer to

ISSN: 2414-1895 DOI: 10.6919/ICJE.202202_8(2).0059

enrich the feature extraction of the network layer.The fast connection performs a simple element-level superposition of the input and output of the block, which can effectively solve the degradation problem in the training process of the deep structure network without adding additional parameters and computation, thus increasing the training speed of the model and improving the training effect. The residual block structure is shown in Figure 2.

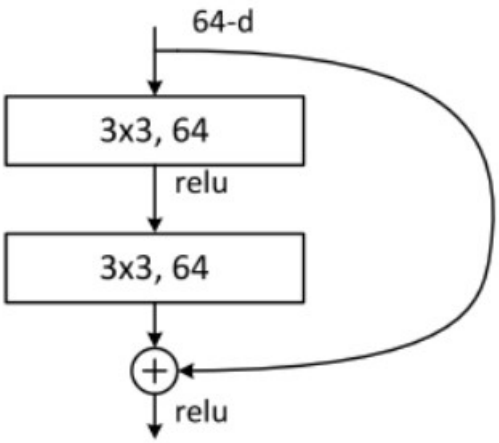

Figure 2. residual block structure

3. Experiment and Ruselts

This paper uses Task 4.3: Word Recognition data set from ICDAR2015 Urban Scene Text.The data set contains many text regions in natural scene images. In the original data, the training set contains 4468 images, and the test set contains 2077 images. All images are generated according to the bounding box of the text region in the original large image, and the text in the image is basically in the center of the image.Adam optimizer is used to automatically adjust the learning rate with the number of iterations.The final experimental accuracy is 92.72 %.

4. Conclusion

Based on ICDAR2015 data set, this paper carries out segmentation and end-to-end recognition based on Transformer model. The whole line of text of variable length in the image is regarded as a text unit to predict the text in the image area and return the content of the text.We plan to further improve the correct recognition rate by starting with parameters.

References

[1] JADERBERG M, SIMONYAN K, VEDALDI A, et al. Reading Text in the Wild with Convolutional Neural Net Works [J]. International Journal of Computer Vision, 2016, 116(1) : 1-20.

[2] LIU X, LIANG D, YAN S, et al. FOTS: Fast Oriented Text Spotting with a Unified Network [J]. 2018.

[3] LUO C, JIN L, SUN Z. MORAN: A Multi-object Rectified attention Network for scene text recognition [J]. ArXiv: 1901. 03003 v1. Jan 2019.

[4] QIN S, ALESSANDRO B, MICHALIS R. Dietspotting [J]. Dietspotting [J]. ArXiv: 1908. 09231 v1, 2019.

[5] SARSHOGH M R, HINES KE. A Multi-task Network for Localization and Recognition of Images [J]. ArXiv: 1906. 09266 v1, 2019.

[6] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]//Proceedings of 2016 IEEEConference on Computer Vision and Pattern Recognition, 2016:770– 778.