ANALYSIS OF THE NOVEL TRANSFORMER MODULE COMBINATION FOR SCENE TEXT RECOGNITION

VASUKI.C

COMPUTER SCIENCE ENGINEERING &

SETHU INSTITUTE OF TECHNOLOGY

Abstract -

In Various methods for scene text recognition (STR) are proposed every year. These methods dramatically increase the performance of the existing STR field; however, they have not been able to keep up with the progress of general-purpose research in image recognition, detection, speech recognition, and text analysis. In this paper, we evaluate the performance of several deep learning schemes for the encoder part of the Transformer in STR. First, we change the baseline feed forward network (FFN) module of encoder to squeeze-andexcitation (SE)-FFN or cross stage partial (CSP)-FFN. Second, the overall architecture of encoder is replaced with local dense synthesizer attention (LDSA) or Conformer structure. Conformer encoder achieves the best test accuracy in various experiments, and SE or CSP-FFN also showed competitive performance when the number of parameters is considered. Visualizing the attention maps from different encoder combinations allows for qualitative performance. Index Terms— Scene Text Recognition (STR), Transformer, Encoder, Self-attention

1.INTRODUCTION

Even though STR is well studied, rapid performance improvement has been achieved recently due to deep learning methods, especially convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Formerly introduced STR papers [2] show how STR field has been developed. Most of the research [3, 4] has been carried out in a serial connection of feature extractor, sequence modeling, and prediction modules, all using CNNs or RNNs. Among these, prediction modules utilize RNN, such as long short-term memory (LSTM) or gated recurrent units (GRU), to perform the attention mechanism. However, there are challenging issues such as irregular shapes, rotations, and noises that are discovered in benchmark test sets or natural images. Recently, Applying the Transformer model instead of RNN-based attention mechanisms is a simple way to handle these issues.

2. Body of Paper

In general, the Transformer architecture consists of three modules: positional encoding, encoder, and decoder. Occasionally, feature extractor are included. Both the feature extractor and the encoder are responsible for embedding the value for the input, and the decoder is responsible for predicting the next character. Specifically, in order to predict the character of the current step, the decoder performs the self-attention with values including output of the encoder and the embedded value of character predicted in the previous step. Thus, the encoder should embed the input image as effectively as possible. In this paper, we test various forms of the Transformer encoders to find the most optimal structure. . Replacing the Feed Forward Network (FFN) This section explains how to transform the baseline-FFN [1] module. Unlike the input of the language model, STR has a two-dimensional (2D) shape input image. Therefore, we figure out that better feature extraction will be possible if the convolution operation is included. Fig. 1 (b) shows a list of proposed FFN modules. 2.2.1. SE Traditional CNN research tends to fuse spatial and channel specific information together within the local receptive field. On the other hand, Hu et al. [11] focused on the relationship between channels and devised SE-Net. We employ SE-block of SE-Net in the FFN module of the Transformer encoder and called it SE-FFN. Among various image recognition architectures [11, 13, 14], SE-block is chosen because it is not too large to be applied to the Transformer and is a radical structure that helps

improve the interdependency problem between channels. Therefore, we can expect not only performance improvement but also additional advantages such as reduction in the number of parameters compared to baseline-FFN.

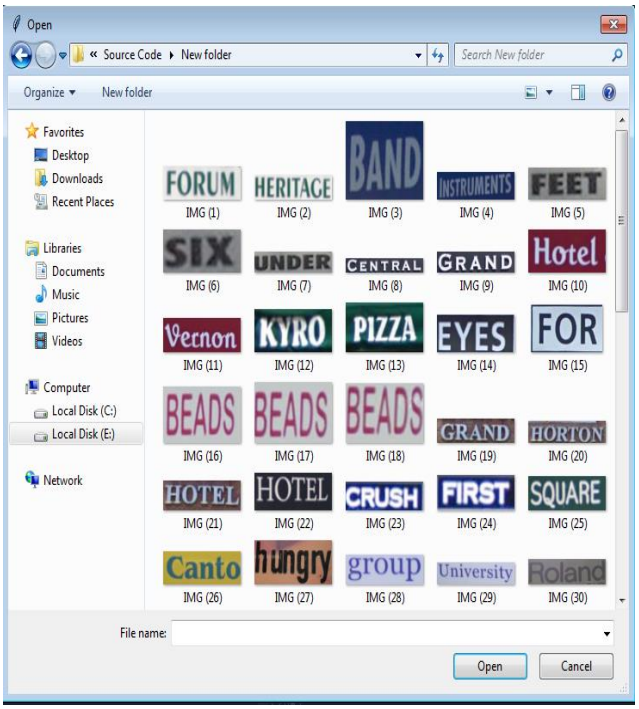
2.2.2. CSP CSP-Net [12] is a new model in the field of image detectors. The concept of CSP-block in CSP-Net is straightforward to apply to the existing deep learning network. CSP-block is similar in principle to the residual neural network (ResNet)-block [15], but it performs convolution operations on only half of the channels. This method improves accuracy while reducing computation cost because the number of gradients to be stored decreases. Therefore, we expect that the CSP-block 1230 Authorized licensed use limited to: IEEE Xplore. Downloaded on August 31,2021 at 20:50:21 UTC from IEEE Xplore. Restrictions apply. can be an effective structure for extracting features of STR images. Thus, we replace baseline-FFN with the CSP-block and called it CSP-FFN. CSP-FFN passes one Rest Net-block for half of the channels and connects it with the rest.

2.3. Replacing the Architecture of Encoder 2.3.1. Conformer Like speech recognition, the network for STR needs to balance both the global and local features of the input image, because even a single word image has a relationship between letters. The first encoder architecture which we propose is a Conformer block as described by Gulati et al. [8] and called Conformer encoder architecture. Specifically, the self-attention of Conformer encoder extracts global features, while the convolution module extracts local features. Additionally, Conformer encoder is an optimized architecture with sandwich-style structure like Macaron-Net [7]. Fig. 2 (a) shows the Conformer encoder applied to the Transformer for STR.

2.3.2. LDSA Transformer includes a self-attention operation that performs matrix multiplications between input tokens. However, it may not be necessary to perform self-attention for all input tokens. Therefore, memory usage and number of computations could be optimized [9, 10]. We apply the LDSA structure, which shows good performance while reducing the amount of computation in the field of speech recognition and called it LDSA encoder architecture. Original Transformer needs complexity of O(n 2) at self-attention module, but LDSA needs only O(nc) where n is the number of input tokens and c is the chunk size of LDSA. We set c as 5. Fig. 2 (b) shows the LDSA encoder applied to the Transformer for STR. The convolution module is the same as [8]. 3. EXPERIMENTS 3.1. Datasets The following three common datasets were used as training datasets. MJSynth consists of 9M synthetic training datasets generated by Jader berg et al. [16]. Synth Text [17] has about 5.5M text bounding boxes all of which are synthetic data; Synth Add has about 1.2M synthetic bounding boxes and are data to reinforce special symbol data generated by [18]. Although the performance can be enhanced by generating and adding training data [19], we use a minimum of training data for fair comparison with other STR papers. we used the following four common test datasets. IIIT5k has 3,000 images collected from the Internet. Street view text (SVT) has 647 images collected from Google Street View. IC03 has 867 images from ICDAR03 robust reading competitions. IC13 has 1,015 images from ICDAR13 focused scene text competitions. In addition, we evaluated on three irregular datasets, which are more challenging due to severe rotation and curves. 2,077 images from IC15 provided by ICDAR2015 incidental scene text competitions were used. 645 images of SVT-Perspective (SVTP), mainly taken side-view in Google Street View, were also used. Finally, 288 images of CUTE80 (CUTE) datasets with severe curve noise were used. Implement Details We set Ne = 12 and Nd = 6, which are the numbers of iterations of the Transformer encoder and decoder. The hidden channel (dmodel) of self-attention is 512, inner dimension (df f) of FFN is 2048 and the number of head (dh) is 8. On the other hands, We set Ne = 9, Nd = 6, and df f = 1024 in Conformer-small architecture. Max length of target sequence is 25 and total 94 characters was trained (52 alphabets, 10 digits, 32 special symbols). Our experiments were trained with Adam optimizer with initial learning rate 0.0003. Our batch size is

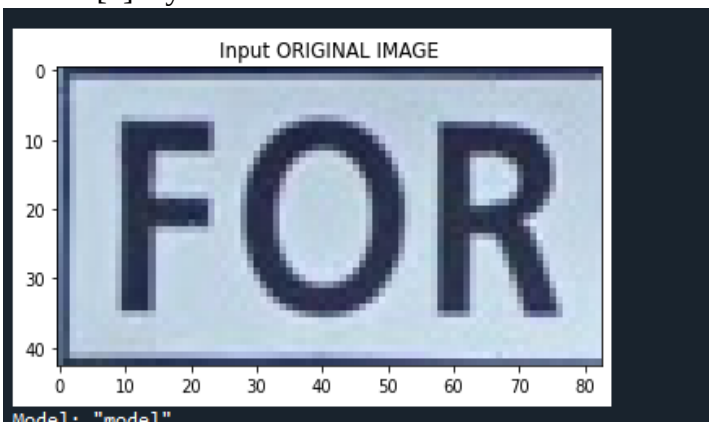
200, and we trained for five epochs. All batches were trained by cross-entropy loss. Additionally, we even tried an extra training experiment that applied data augmentation [20] to our best model.

Evaluation We varied the encoder architecture and FFN module and present the results in Table 1. The number of parameters for both CSP-FFN and SE-FFN decreases while the test accuracy increases. A baseline-FFN module requires 2M parameters with inner dimension (df f) of 2048 between two linear layers, but each single SE-FFN and CSP-FFN module demand only 800K and 1.2M respectively. The reductions in SE-FFN and CSP-FFN are induced by decreasing the number of internal channels, while all connection weights of neurons are calculated in the linear function. Although the number of parameters was reduced, the accuracy from seven test datasets was improved or maintained, except for the CUTE accuracy in CSP-FFN. In the case of Conformer-small encoder, the number of parameters was reduced over the baseline, and the test accuracy was improved on five test datasets compared to baseline encoder. Conformer-big encoder increases the number of parameters but improves the test accuracy for all test datasets. This outcome might be explained as a well-balanced extraction of local features and global features. Also, FFN module placing at the front and the end, like the Macaron-Net [7], works effectively. Test accuracy of the LDSA encoder was like that of the baseline encoder; however, the number of parameters increased as the convolution module was included. The test accuracy comparison with other STR models is presented in Table 2. Three new performance records (IIIT5k, SVT, SVT P) and one tie (IC15) are achieved. The visualized attention maps according to FFN module are shown in Fig. 3. SE-FFN and CSP-FFN both show more elaborate attention map distributions than baseline-FFN with the attention drift problem. The difference in the distribution of the attention maps according to LDSA and Conformer encoder is observed in Fig. 3 as well. In the case of LDSA encoder, the attention map spans several letters. This seems to refer strongly to the surrounding characters as much as the LDSA encoder chunk size of five. For Conformer encoder, the range of the attention map is narrow, but it appears more accurate than any other architecture or modules. Because the added convolution module in Conformer encoder emphasizes local features, attention maps are narrowed. In general, the self-attention operation of the Transformer emphasizes the global features of the input, and convolution operation emphasizes the local features of the input. Thus, it seems that the global and local features are harmoniously emphasized in the Conformer encoder architecture.



CONCLUSIONS

In this paper, several novel Transformer architectures are proposed in STR with extensive experiments. Conformer encoder achieved record performance, and SE-FFN and CSPFFN showed sufficiently competitive performance considering the number of parameters. In addition, the possibility of improving the efficiency of self-attention in STR field such as LDSA was confirmed. These results prove that the convolution operation and Macaron-Net [7] style is valid for the Transformer when the input has 2D shape, especially in STR. Furthermore,



our attention maps prove that the performance was properly improved by solving the attention drift problem.

ACKNOWLEDGEMENT
The heading should be treated as a 3rd level heading and should not be assigned a number.

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, pp. 5998–6008, 2017.
[2] Xiaoxue Chen, Lianwen Jin, Yuanzhi Zhu, Canjie Luo, and Tianwei Wang, "Text recognition in the wild: A survey," ACM Computing Surveys (CSUR), vol. 54, no. 2, pp. 1–35, 2021.
[3] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai, "Decoupled attention network for text recognition.," in AAAI, 2020, pp. 12216–12224.
[4] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4715–4723.