Scene Text Recognition with Self-supervised Contrastive Predictive Coding

Xinzhe Jiang[1], Jianshu Zhang[2], Jun Du[1,*], Zhenrong Zhang[1], Jiajia Wu[2]
[1]National Engineering Research Center of Speech and Language Information Processing
University of Science and Technology of China, Hefei, Anhui, P. R. China
[2]iFLYTEK Research
Email: xzjiang@mail.ustc.edu.cn, jszhang6@iflytek.com
jundu@ustc.edu.cn, zzr666@mail.ustc.edu.cn, jjwu@iflytek.com

***Abstract*—Self-supervised visual pre-training has recently emerged in scene text recognition (STR), which designs the pretext tasks and takes unlabeled data as input to obtain useful representations for STR. However, most current self-supervised methods do not pay special attention to the importance of sequence awareness. Accordingly, we propose a novel self-supervised STR method based on contrastive predictive coding (STR-CPC), which regards a text instance as a sequence from left to right and captures the visual sequence correlation. Considering the information overlap problem within the feature map induced by the deep convolutional neural network (CNN) encoder, we design a widthwise causal convolution during model pre-training and a progressive recovery training strategy (PRTS) during model fine-tuning to improve the STR performance. Experiments on scene text show that our STR-CPC method outperforms the existing self-supervised methods, which testifies the advantage of visual sequence correlation for STR. Additionally, STR-CPC observably boosts performance compared with supervised training when the amount of labeled data decreases.**

I. Introduction

Text recognition [1]–[5] attempts to translate the text instance into machine-readable text, which plays a crucial role in daily applications such as autonomous vehicles navigation [6] and document electronization [7].

There is a consensus that the mainstream models for scene text recognition (STR) are trained on large synthetic data such as MJSynth (MJ) [8] and SynthText (ST) [9], instead of real-world text images. However, [4] claims that the synthetic data lacks diversity which is more important than the amount of data. The real-world labeled data can bring more diversity into STR, while its annotation is time-consuming and expensive. Hence, it is cost-effective to explore the hidden potential of the large-scale unlabeled data via self-supervised representation learning without human annotation.

The self-supervised learning frameworks formulate pretext tasks and help models learn useful representations, which have achieved outstanding success in the field of image classification [10]–[12]. Nevertheless, self-supervised learning for text recognition has rarely been researched. [13] utilizes two self-supervised methods for STR, RotNet [14] and MoCo [11]. The two methods treat the text instance as a whole image rather than a semantic sequence, then undertake the rotation prediction and instance discrimination tasks respectively, which lack the sequence awareness for text instances. [15] proposes a framework of sequence-to-sequence contrastive learning (SeqCLR), applied to the individual elements of the sequence using an instance-mapping function, which aims at instance discrimination and cares less about the visual sequence correlation within the text instances.

Where the sequence correlation is concerned, wav2vec [16], which utilizes the contrastive predictive coding (CPC) with a contrastive loss, takes raw audio as input, and extracts slow features that can be fed to a speech recognition system. However, when a text instance is regarded as a sequence from left to right, the vanilla CPC [17] cannot perform well in STR due to the information overlap problem. A deep convolutional neural network (CNN) with a large receptive field is often adopted in a traditional STR model, making the CPC task trivial since future information has already been leaked during the encoding process.

In this paper, we introduce a self-supervised STR method based on contrastive predictive coding (STR-CPC), exploiting low-cost unlabeled text instances to capture the visual sequence correlation. STR-CPC possesses a CNN encoder that takes text instance images as input and generates useful representations that can be further fed to the downstream STR models. We also utilize the InfoNCE [17] loss as the training objective of STR-CPC, which requires the model to distinguish a positive future sample from negatives. Moreover, we design a widthwise causal convolution during pre-training to alleviate the information overlap problem and then introduce a progressive recovery training strategy (PRTS) that progressively converts the widthwise causal convolution to the common convolution for improving the STR performance during fine-tuning.

The main contributions of our work are summarized as:

- We introduce the self-supervised STR-CPC for scene text recognition and explore the information overlap problem in the vanilla CPC.
- We design the widthwise causal convolution and progressive recovery training strategy to mitigate the limitation of the information overlap problem.
- Extensive experiments show that the proposed STR-CPC explicitly improves the performance of the STR models and outperforms the existing self-supervised methods.

[*] Corresponding Author

When the amount of labeled data decreases, STR-CPC observably boosts performance compared with supervised training, indicating its capability in low-resource settings.

To our best knowledge, STR-CPC is the first predictive-coding-based self-supervised method for text recognition, and it further leads to significant improvement for the STR models.

II. Related work

A. Scene Text Recognition

Scene text recognition (STR) has received extensive attention due to its universal application in daily life. The mainstream STR models can be divided into two approaches according to their decoding manners, Connectionist Temporal Classification (CTC) [18] based ones and attention mechanism based ones. For CTC-based methods, [1], [19] adopt LSTM for sequence modeling and CTC for prediction. These methods are effective and practical for real-world application scenarios. [20] gives a theoretical explanation of CTC from the viewpoint of the Expectation-Maximization algorithm and proposes a pseudo-label-based L1 regularization and voting decoding algorithm to improve the performance of text recognition. For attention-based methods, [2] proposes an attention and language ensemble method to boost prediction jointly, with both visual cues and linguistic rules captured. [3] proposes a sequential transformation network consisting of a series of patch-wise basic transformations, which makes the irregular scene text images more readable for the attention-based recognizer.

B. Unsupervised Learning for Text Recognition

In spite of the success of unsupervised representation learning in computer vision tasks such as image classification and object detection [10]–[12], most text recognition methods have not taken advantage of enormous unlabeled text images.

[21] proposes an unsupervised method that learns a predictor to convert images into strings that statistically match the target corpora, implicitly reproducing quantities such as letter and word frequencies and n-grams. [15] proposes a framework of sequence-to-sequence contrastive learning (SeqCLR) and divides each feature map into different instances over which the contrastive loss is computed. [13] uses two self-supervised methods, RotNet [14] and MoCo [11], which consider the input text instance as a whole to accomplish the rotation prediction and instance discrimination tasks respectively.

C. Contrastive Predictive Coding

There are a number of works utilizing a predictive-coding-based method called contrastive predictive coding (CPC) to learn useful latent representations from unlabeled data in an unsupervised manner. The reason why using a contrastive objective rather than a generative one is that the generative objective tends to make the model focus on minor details and local noise [22]. [17] compresses high-dimensional data into a compact latent embedding space and utilizes Noise-Contrastive Estimation [23] for the loss function. [16] applies unsupervised pre-training to improve supervised speech recognition, adopting a contrastive loss that requires the model to distinguish a true future audio sample from negatives. [24] modifies CPC via patch-based image augmentations and predictions in four directions, improving the ImageNet classification accuracy.

III. Proposed approach

As the proposed self-supervised approach for STR is based on CPC, we first review this algorithm concisely and then discuss why the vanilla CPC cannot perform well for STR. Subsequently, we detail our STR-CPC framework and the downstream STR model fine-tuning.

A. Preliminary

1) Vanilla CPC: CPC forces the model to predict future timesteps in the latent space. With the text instances regarded as sequences, the left and right within text instances correspond to the past and future in time. We utilize the vanilla CPC method to process the text instances from left to right, predicting the future based on past information.

A non-linear encoder network is utilized to map an unlabeled text instance image x to a sequence of latent representations z_t. Subsequently, a context network is used to provide a context latent representation c_t. Instead of generating the future samples x_{t+k} directly via a generative model, an unnormalized density ratio $f_k(x_{t+k}, c_t)$ which estimates the mutual information between x_{t+k} and c_t, is calculated by a log-bilinear model:

$$f_k\left(x_{t+k}, c_t\right) = \exp\left(z_{t+k}^{\top} W_k c_t\right) \tag{1}$$

A linear transformation $W_k c_t$ is adopted to predict the k future timesteps, with a separate W_k corresponding to each step k. The density ratio $f_k(x_{t+k}, c_t)$ calculates the similarity between z_{t+k} and c_t, where the similarity score means the prediction probability. The model parameters are updated via optimizing the InfoNCE loss function:

$$\mathcal{L} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k\left(x_{t+k}, c_t\right)}{\sum_{x_j \in X} f_k\left(x_j, c_t\right)}\right] \tag{2}$$

The InfoNCE loss is intrinsically the categorical cross-entropy, classifying the positive sample correctly from a set $X = \{x_1, \ldots, x_N\}$. Note that the set X consists of N random samples, in which one positive sample is from $p\left(x_{t+k}|c_t\right)$ and $N-1$ negative samples are from the 'proposal' distribution $p\left(x_{t+k}\right)$. To be specific, the negative samples are obtained by uniformly choosing distractors from each feature map itself. For the context vector c_t, $\{z_{t+1}, \ldots z_{t+k}\}$ are the positive samples.

$$I\left(x_{t+k}, c_t\right) \geq \log(N) - \mathcal{L} \tag{3}$$

The mutual information between the c_t and x_{t+k} can be evaluated as Equation 3, which points out that a lower bound on mutual information is maximized via optimizing the InfoNCE loss function.

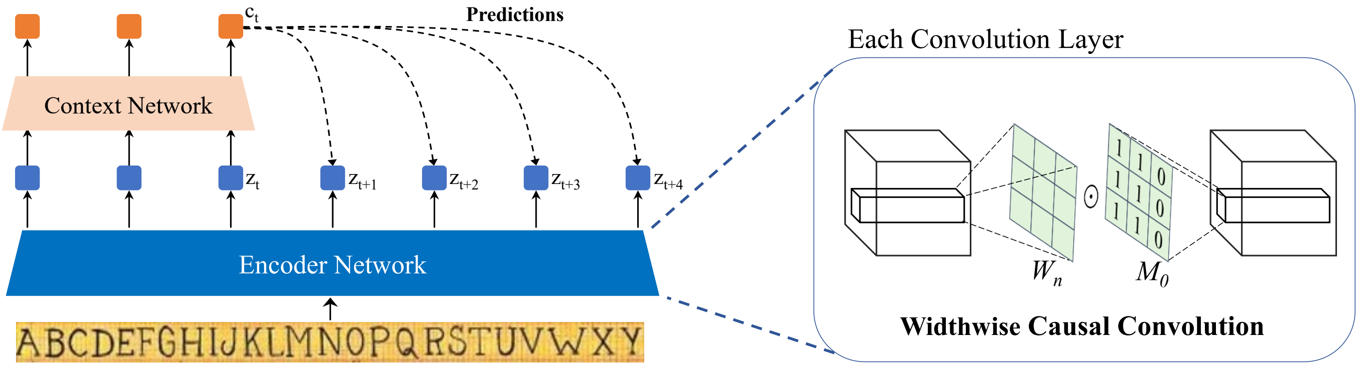

Fig. 1: The architecture of self-supervised STR-CPC pre-training. The encoder and context network are the WC-CNN and linear projection, respectively. The model predicts the future time steps based on the past context.

2) Why Vanilla CPC Cannot Perform Well for STR: Different from time-series data like audio signals, the nonlinear encoder of text instance images always prefers the deeper CNN. The convolution operation for the input text instance images results in the information overlap problem within adjacent timesteps in the feature map. Hence the pretext classification task is likely to become trivial when it predicts the adjacent future timesteps. The deeper the CNN encoder is, the larger the receptive field size is, and the more trivial the pretext task will be.

Equation 4 is the objective function under the ideal condition, which means there is no information overlap problem.

$$\mathcal{L} = -\mathbb{E}_X \left[\frac{z_{t+k}^{\top} W_k c_t}{\sum\limits_{\tilde{z} \sim p_n} \tilde{z}^{\top} W_k c_t} \right] \tag{4}$$

where $\tilde{z}$ is the negative sample drawn from p_n. In practice, we set the negative sampling distribution as $p_n(z) = \frac{1}{T}$, where T is the width of the feature map z.

Considering the information overlap problem induced by the large receptive field size of the encoder, there are many overlaps between the past and future. On this account, the information of the future z_{t+k} leaks into the past context c_t, and predicting the nearby future timesteps becomes trivial. The problem seriously hinders the model training and leads to incorrect convergence. The problem induces the worse lower bound on mutual information, and the pre-trained encoder cannot provide high-quality representations for STR. The main challenge of applying CPC in STR is the information overlap problem, which motivates us to propose the STR-CPC method.

B. Our STR-CPC Approach

In this section, we introduce the proposed STR-CPC framework, which includes: (1) STR-CPC pre-training as shown in Fig. 1, where unlabeled text instances are exploited to obtain effective representations via the pretext task; and (2) STR model fine-tuning as shown in Fig. 2, which integrates the pre-trained encoder into the downstream STR models for feature extraction.

1) Widthwise Causal Convolution for STR-CPC Pre-training: The proposed STR-CPC takes unlabeled text instances as input and considers them as the sequences from left to right. STR-CPC enables the model to extract high-level information and incorporates the visual sequence correlation into latent representation learning. We hypothesize that the visual sequence correlation within text instances is beneficial to learning effective representations for STR.

The architecture for STR-CPC is depicted in Fig. 1. Inspired by the causal convolutional layers in [25], we design a widthwise causal convolution to decrease information overlaps between the past and the future, where the right half of the kernel in convolution is masked by zero. The widthwise causal CNN (WC-CNN) encoder consists of widthwise causal convolution layers, while the common CNN encoder consists of common convolution layers without mask operation. The widthwise causal convolution is defined as Equation 5.

$$g_{\text{wcc}}(x) = b + \sum_{n=1}^{D} M_0 \cdot W_n \star x_n \tag{5}$$

where x_n is the n-th channel of the input x, D is the number of input channels, b is the bias term, W_n is the kernel weights matrix for x_n, M_0 is the widthwise causal mask for W_n, $\cdot$ is the Hadamard product, $\star$ is the valid 2D cross-correlation operator. If the widthwise causal mask M_0 is removed, the widthwise causal convolution will revert to the common convolution.

According to [26], we can compute the receptive field size of the L-layer CNN by Equation 6.

$$r_o = \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + 1 \tag{6}$$

where k_l and s_l are the kernel size and the stride of l-th layer, and r_o is the receptive field size.

Once the original common convolution is substituted by widthwise causal convolution, the width of kernel for l-th layer is reduced from k_l to $\lceil \frac{k_l}{2} \rceil$. When we only consider the convolution layers, the receptive field size of the WC-CNN encoder

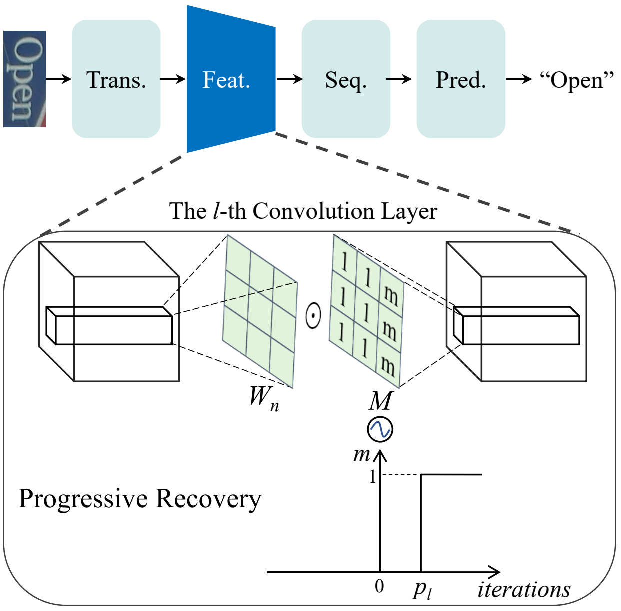

Fig. 2: The STR model with PRTS during fine-tuning. The typical STR models consist of four modules: transformation (Trans.), feature extraction (Feat.), sequence modeling (Seq.), and Prediction (Pred.). The l-th convolution layer will convert to common convolution from widthwise causal convolution when the training iterations reach p_l.

declines by $\Delta r = \sum_{l=1}^{L}\left(\lfloor \frac{k_l}{2} \rfloor \prod_{i=1}^{l-1} s_i\right)$ compared with the common CNN encoder. Consequently, the information overlap problem is greatly alleviated.

2) Progressive Recovery Training Strategy for STR Model Fine-tuning: Given the disparity between the WC-CNN encoder and the common CNN encoder, there are structural mismatches between pre-training and fine-tuning. The mismatches inevitably degenerate model performance during fine-tuning. As illustrated in Fig. 2, we propose the progressive recovery training strategy (PRTS) to reduce mismatches via progressive switching from widthwise causal convolution to common convolution in fine-tuning. The training strategy is defined as Equation 7 and 8.

$$g_l\left(x\right) = b + \sum_{n=1}^{D} M \cdot W_n \star x_n \tag{7}$$

$$m_{i,j} = \begin{cases} 1 & j \leqslant \lceil \frac{k_l}{2} \rceil \\ u(iterations - p_l) & j > \lceil \frac{k_l}{2} \rceil \end{cases} \tag{8}$$

where g_l is the l-th convolution layer, M is the mask matrix for W_n, $m_{i,j}$ is the element of row i and column j of M. $u\left(\cdot\right)$ is the unit step function, and $p_1 > p_2 > \cdots > p_l$. As *iterations* reaches p_l, the g_l is converted to the common convolution from widthwise causal convolution.

At the beginning of fine-tuning, the feature extractor of the STR models is defined as the WC-CNN encoder and initialized from the STR-CPC pre-trained WC-CNN encoder weights. As training iteration goes, the widthwise causal convolution is progressively converted to common convolution layerwisely, which preserves the latent visual sequence correlation and improves the STR model performance.

IV. Experiments

A. Datasets

Three public datasets Book32 [27], TextVQA [28], and ST-VQA [29] are pre-processed and consolidated by [13] for self-supervised pre-training, which is called the Real-U dataset. With 4.2M unlabeled text instances obtained, we utilize the Real-U dataset for the STR-CPC pre-training in our experiments. During fine-tuning, we use the Real-L dataset for STR model training, same as [13], which includes 276K training and 63K validation sets.

Six standard benchmark datasets include SVT [30], IIIT [31], IC13 [32], IC15 [33], SVTP [34] and CUTE [35] are adopted to evaluate the performance of STR models. The word-level accuracy is calculated only on the alphabet and digits, in accordance with [36]. For convenience, we calculate the total accuracy of the union of six benchmark datasets (7,672 in total). In the following, the accuracy indicates total accuracy for performance comparison.

B. STR Baseline Models

According to [4], STR is performed in four stages, which include transformation (Trans.), feature extraction (Feat.), sequence modeling (Seq.), and prediction (Pred.) as shown in Fig. 2. We specifically adopt two classic STR models: CRNN [1] and TRBA [4]. To ensure a fair comparison, the chosen STR models are the same as [13]. CRNN is a CTC-based STR model without the transformation stage, which consists of VGG, BiLSTM, and CTC. Concerned with performance, CRNN is inferior to state-of-the-art methods, but it is widely used in practical applications due to its fast speed and low memory requirement. TRBA is a typical attention-based STR model consisting of TPS [37], ResNet, BiLSTM, and Attention for each stage, which has higher accuracy than CRNN, with relatively slow speed and high memory requirement.

C. Implementation Details

1) STR-CPC Pre-training: We use the Real-U dataset for STR-CPC pre-training. Specifically, we choose the VGG and ResNet as the encoder networks, corresponding to the feature extractors of CRNN and TRBA respectively. The prediction networks W_k are linear layers, and the prediction step k is 8. All pre-training experiments are implemented with 4 NVIDIA P40 GPUs.

2) STR Model Fine-tuning: In all our experiments, we use the Real-L dataset during fine-tuning stage. We adopt the Adam [38] optimizer and the one-cycle learning rate scheduler [39] with a maximum learning rate of 0.0005, and train the STR models on a single NVIDIA P40 GPU.

TABLE I: Accuracy of STR models on six benchmark datasets for comparison. Methods with symbol $*$ denote the results are reported while the others are our implementations. CRNN† means the model setting is different from our CRNN.

Model	Pre-training method	Training data	IIIT	SVT	IC13	IC15	SVTP	CUTE	Total	Total$_{*0.1}$
CRNN†	SeqCLR [15]*	ST [9]	80.9	-	86.3	-	-	-	-	-
TRBA	SeqCLR*	ST	82.9	-	87.9	-	-	-	-	-
CRNN	None	Real-L	83.5	75.5	86.3	62.2	60.9	64.7	74.8	40.9
	MoCo [11]	Real-L	83.9	77.9	85.9	62.9	63.1	65.9	75.6	56.9
	RotNet [14]	Real-L	84.5	77.4	87.5	64.2	63.3	64.9	76.3	56.9
	Vanilla CPC [17]	Real-L	83.8	77.6	85.4	62.0	61.1	64.6	75.0	54.3
	STR-CPC w/o PRTS	Real-L	82.7	78.1	84.8	60.9	61.9	63.5	74.2	56.1
	STR-CPC	Real-L	84.7	79.6	88.7	64.7	64.3	68.1	**77.0**	**58.7**
TRBA	None	Real-L	93.5	87.5	92.6	76.0	78.7	86.1	86.6	41.3
	MoCo	Real-L	92.2	86.2	91.8	73.2	76.6	83.7	84.9	59.8
	RotNet	Real-L	92.8	87.2	92.0	75.7	77.8	85.0	85.8	62.2
	Vanilla CPC	Real-L	92.9	88.1	91.9	75.5	77.1	87.2	86.1	61.9
	STR-CPC w/o PRTS	Real-L	93.3	88.4	92.5	75.0	80.3	87.9	86.5	64.7
	STR-CPC	Real-L	93.4	88.1	93.6	76.9	79.5	88.2	**87.2**	**66.9**

TABLE II: Ablation study on the number of convolution layers and kernel size of VGG in CRNN.

VGG Model Settings		Pre-training method		
		None	Vanilla CPC	STR-CPC
Convolution Layers	7	40.9	54.3	58.7
	9	43.5	49.5	58.3
	11	47.3	51.6	59.9
Kernel Size	3 × 3	40.9	54.3	58.7
	5 × 5	41.0	48.7	53.8
	7 × 7	33.2	39.3	47.0

D. STR Performance

The comparison of our STR-CPC method and other self-supervised pre-training methods is shown in Table I, where 'Total' is the total accuracy and 'Total$_{*0.1}$' is the total accuracy when only 10% labeled data is used for training. In Table I, the 'None' in the column 'Pre-training method' refers to the baseline STR models, which are trained in a supervised manner without any self-supervised pre-training. To strictly perform a fair comparison, we reproduce two self-supervised methods, MoCo and RotNet. In the aspect of SeqCLR, the CRNN and training data are not exactly the same as ours. Therefore, we do not focus on comparing the results between the STR-CPC and SeqCLR.

In the aspect of CRNN, MoCo, RotNet, and STR-CPC methods improve the accuracy by 0.8%, 1.5%, and 2.2% from CRNN baseline model. Although MoCo and RotNet methods achieve improvements for CRNN on benchmarks, STR-CPC gets better performance compared with them. For TRBA, the STR-CPC improves the accuracy by 0.6% from TRBA baseline model. Contrary to CRNN, TRBA model performance is even degraded by the MoCo and RotNet pre-training methods.

To evaluate the influence of the information overlap problem, we conduct the ablation study on the receptive field size of VGG in CRNN with 10% training data. We enlarge the receptive field size of VGG by utilizing the more convolution layers and larger kernel size, respectively. Note that we do not simultaneously change the number of convolution layers and kernel size. The basic VGG consists of seven convolution layers and 3 × 3 kernels. Table II shows that the performance increment of vanilla CPC is degraded gradually with the increase of convolution layers and kernel size. However, the proposed STR-CPC tends to keep substantial improvement upon the baseline model performance when the receptive field size of VGG is enlarged.

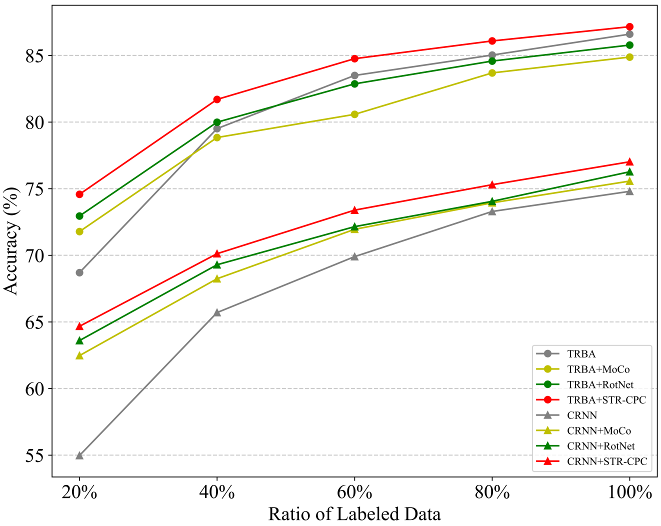

Fig. 3: Accuracy of different methods with the decrease of labeled data.

For both CRNN and TRBA, our STR-CPC outperforms the vanilla CPC. The vanilla CPC has a more negative influence on TRBA than CRNN due to the deeper CNN encoder and larger receptive field size in TRBA. As can be seen from the comparison in Table I, the removal of PRTS decreases the accuracy by 2.8% and 0.7% for the STR-CPC pre-trained CRNN and TRBA models, which indicates that the combination of widthwise causal convolution and PRTS is valuable to the STR-CPC. In the scenarios of all labeled data and 10% labeled data for training, the STR-CPC achieves the best results on STR benchmarks for both CRNN and TRBA models, which demonstrates the capability of the STR-CPC.

E. Varying Amount of Labeled Data

To explore the effectiveness of self-supervised pre-training methods in low-resource settings, we decrease the amount of labeled data proportionately and conduct experiments for comparison. Note that the amount of unlabeled data of pre-training keeps unchanged. As shown in Fig. 3, the accuracies of CRNN and TRBA baseline models descend 19.8% and 17.9% respectively when the ratio of labeled data is reduced from 100% to 20%, which demonstrates the baseline models cannot perform well in low-resource settings. Moreover, the STR-CPC observably boosts the performance (+9.7% for CRNN and +5.9% for TRBA) and excels the RotNet and MoCo methods when only 20% labeled data is utilized during fine-tuning. When the amount of labeled data proportionately rises, the performance of all methods increases, whereas the gains of self-supervised methods decline.

F. Representation Quality

To further compare the representation quality of different self-supervised methods, we train the STR models with all labeled data during fine-tuning and keep the pre-trained encoders frozen. As illustrated in Fig. 4, the accuracy comparison among different self-supervised methods demonstrates that the best result is accomplished by our STR-CPC method. In addition, it is worth noting that the RotNet method can provide few high-quality representations for STR when the pre-trained encoders are frozen during fine-tuning, even though it gains an increment for CRNN in low-resource settings as shown in Fig. 3.

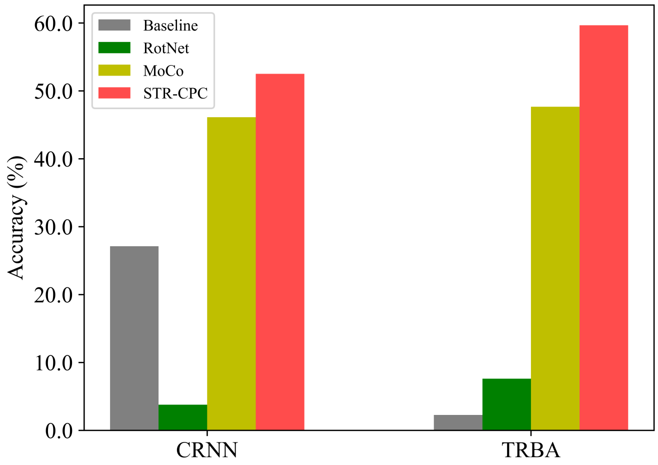

Fig. 4: Accuracy of different methods with frozen encoders.

G. Performance Comparison of Different Text Lengths

Due to the error accumulation among decoding steps, the attention-based models tend to perform poorly for the long text instances. The proposed STR-CPC method equipped with sequence awareness can strengthen the attention-based TRBA model performance in the situation of long text instances. For a clearer comparison, we divide the STR benchmark datasets into three groups according to the label lengths and calculate the Word Error Rate (WER) for each group. As shown in Fig. 5, experiment results demonstrate that the STR-CPC outperforms other self-supervised methods for the long text instances. For groups 5-8 and 9-12, STR-CPC achieves 2.0% and 2.9% reductions on WER for TRBA baseline model, respectively. Benefiting from STR-CPC pre-training, TRBA model is able to capture the visual sequence correlation and ease the restriction of error accumulation among decoding steps.

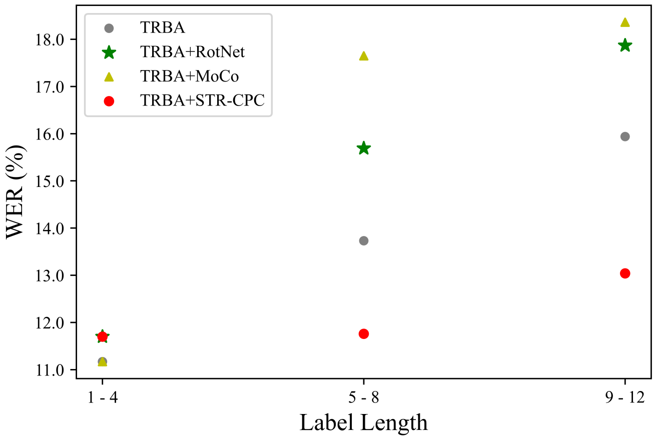

Fig. 5: The Word Error Rate of different methods for variable label lengths. For instance, group 1-4 includes the samples whose label lengths range from 1 to 4.

V. Conclusion

In this paper, we propose a self-supervised STR method based on Contrastive Predictive Coding (STR-CPC), which regards a text instance as a sequence from left to right and forces the model to predict future timesteps in the latent space. The combination of the widthwise causal convolution in pre-training and the PRTS in fine-tuning is capable of mitigating the information overlap problem and improving the performance of STR models. The key takeaway is that the latent visual sequence correlation within the text instance is beneficial to the performance of the STR models. Experiments on STR benchmarks demonstrate that the proposed STR-CPC method not only boosts the performance of the baseline models but also outperforms the existing self-supervised methods. Moreover, with the decrease of labeled data, STR-CPC considerably improves model performance compared with supervised training. In the future, we will further explore the performance of STR-CPC for other languages like Chinese to validate its generalization capability.

Nowadays, the self-supervised works for STR are still limited. We hope that our STR-CPC method can inspire other researchers to offer deeper insight into self-supervised learning for text recognition in the future.

References

[1] B. Shi, X. Bai, and C. Yao, "An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2298–2304, 2017. [Online]. Available: https://doi.org/10.1109/TPAMI.2016.2646371

[2] S. Fang, H. Xie, Z. Zha, N. Sun, J. Tan, and Y. Zhang, "Attention and language ensemble for scene text recognition with convolutional sequence modeling," in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, S. Boll, K. M. Lee, J. Luo, W. Zhu, H. Byun, C. W. Chen, R. Lienhart, and T. Mei, Eds. ACM, 2018, pp. 248–256. [Online]. Available: https://doi.org/10.1145/3240508.3240571

[3] Q. Lin, C. Luo, L. Jin, and S. Lai, "STAN: A sequential transformation attention-based network for scene text recognition," *Pattern Recognit.*, vol. 111, p. 107692, 2021. [Online]. Available: https://doi.org/10.1016/j.patcog.2020.107692

[4] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, "What is wrong with scene text recognition model comparisons? dataset and model analysis," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 4714–4722. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00481

[5] J. Zhang, Y. Zhu, J. Du, and L. Dai, "Radical analysis network for zero-shot learning in printed chinese character recognition," in *2018 IEEE International Conference on Multimedia and Expo, ICME 2018, San Diego, CA, USA, July 23-27, 2018*. IEEE Computer Society, 2018, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICME.2018.8486456

[6] Á. Gonzalez, L. M. Bergasa, and J. J. Y. Torres, "Text detection and recognition on traffic panels from street-level imagery using visual appearance," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 1, pp. 228–238, 2014. [Online]. Available: https://doi.org/10.1109/TITS.2013.2277662

[7] S. Long, X. He, and C. Yao, "Scene text detection and recognition: The deep learning era," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 161–184, 2021. [Online]. Available: https://doi.org/10.1007/s11263-020-01369-0

[8] M. Jaderberg, K. Simonyan, A. Vedaldi, and A. Zisserman, "Synthetic data and artificial neural networks for natural scene text recognition," *CoRR*, vol. abs/1406.2227, 2014. [Online]. Available: http://arxiv.org/abs/1406.2227

[9] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 2315–2324. [Online]. Available: https://doi.org/10.1109/CVPR.2016.254

[10] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bklr3j0cKX

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshick, "Momentum contrast for unsupervised visual representation learning," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 9726–9735. [Online]. Available: https://doi.org/10.1109/CVPR42600.2020.00975

[12] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 1597–1607. [Online]. Available: http://proceedings.mlr.press/v119/chen20j.html

[13] J. Baek, Y. Matsui, and K. Aizawa, "What if we only use real datasets for scene text recognition? toward scene text recognition with fewer labels," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 3113–3122. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Baek_What_if_We_Only_Use_Real_Datasets_for_Scene_Text_CVPR_2021_paper.html

[14] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=S1v4N2l0-

[15] A. Aberdam, R. Litman, S. Tsiper, O. Anschel, R. Slossberg, S. Mazor, R. Manmatha, and P. Perona, "Sequence-to-sequence contrastive learning for text recognition," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 15 302–15 312. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2021/html/Aberdam_Sequence-to-Sequence_Contrastive_Learning_for_Text_Recognition_CVPR_2021_paper.html

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 3465–3469. [Online]. Available: https://doi.org/10.21437/Interspeech.2019-1873

[17] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018. [Online]. Available: http://arxiv.org/abs/1807.03748

[18] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, ser. ACM International Conference Proceeding Series, W. W. Cohen and A. W. Moore, Eds., vol. 148. ACM, 2006, pp. 369–376. [Online]. Available: https://doi.org/10.1145/1143844.1143891

[19] P. He, W. Huang, Y. Qiao, C. C. Loy, and X. Tang, "Reading scene text in deep convolutional sequences," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, D. Schuurmans and M. P. Wellman, Eds. AAAI Press, 2016, pp. 3501–3508. [Online]. Available: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12256

[20] L. Gao, H. Zhang, and C. Liu, "Regularizing CTC in expectation-maximization framework with application to handwritten text recognition," in *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*. IEEE, 2021, pp. 1–7. [Online]. Available: https://doi.org/10.1109/IJCNN52387.2021.9533713

[21] A. Gupta, A. Vedaldi, and A. Zisserman, "Learning to read by spelling: Towards unsupervised text recognition," in *ICVGIP 2018: 11th Indian Conference on Computer Vision, Graphics and Image Processing, Hyderabad, India, 18-22 December, 2018*. ACM, 2018, pp. 33:1–33:10. [Online]. Available: https://doi.org/10.1145/3293353.3293386

[22] B. Millidge, A. K. Seth, and C. L. Buckley, "Predictive coding: a theoretical and experimental review," *CoRR*, vol. abs/2107.12979, 2021. [Online]. Available: https://arxiv.org/abs/2107.12979

[23] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, ser. JMLR Proceedings, Y. W. Teh and D. M. Titterington, Eds., vol. 9. JMLR.org, 2010, pp. 297–304. [Online]. Available: http://proceedings.mlr.press/v9/gutmann10a.html

[24] O. J. Hénaff, "Data-efficient image recognition with contrastive predictive coding," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 4182–4192. [Online]. Available: http://proceedings.mlr.press/v119/henaff20a.html

[25] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*. ISCA, 2016, p. 125. [Online]. Available: http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_DS-4_van_den_Oord.html

[26] A. Araujo, W. Norris, and J. Sim, "Computing receptive fields of convolutional neural networks," *Distill*, vol. 4, no. 11, p. e21, 2019.

[27] B. K. Iwana and S. Uchida, "Judging a book by its cover," *CoRR*, vol. abs/1610.09204, 2016. [Online]. Available: http://arxiv.org/abs/1610.09204

[28] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards VQA models that can read," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer

Vision Foundation / IEEE, 2019, pp. 8317–8326. [Online]. Available: http://openaccess.thecvf.com/content_CVPR_2019/html/Singh_Towards_VQA_Models_That_Can_Read_CVPR_2019_paper.html

[29] A. F. Biten, R. Tito, A. Mafla, L. G. i Bigorda, M. Rusiñol, C. V. Jawahar, E. Valveny, and D. Karatzas, "Scene text visual question answering," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 2019, pp. 4290–4300. [Online]. Available: https://doi.org/10.1109/ICCV.2019.00439

[30] K. Wang, B. Babenko, and S. J. Belongie, "End-to-end scene text recognition," in *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, D. N. Metaxas, L. Quan, A. Sanfeliu, and L. V. Gool, Eds. IEEE Computer Society, 2011, pp. 1457–1464. [Online]. Available: https://doi.org/10.1109/ICCV.2011.6126402

[31] A. Mishra, K. Alahari, and C. V. Jawahar, "Scene text recognition using higher order language priors," in *British Machine Vision Conference, BMVC 2012, Surrey, UK, September 3-7, 2012*, R. Bowden, J. P. Collomosse, and K. Mikolajczyk, Eds. BMVA Press, 2012, pp. 1–11. [Online]. Available: https://doi.org/10.5244/C.26.127

[32] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. Almazán, and L. de las Heras, "ICDAR 2013 robust reading competition," in *12th International Conference on Document Analysis and Recognition, ICDAR 2013, Washington, DC, USA, August 25-28, 2013*. IEEE Computer Society, 2013, pp. 1484–1493. [Online]. Available: https://doi.org/10.1109/ICDAR.2013.221

[33] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. K. Ghosh, A. D. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "ICDAR 2015 competition on robust reading," in *13th International Conference on Document Analysis and Recognition, ICDAR 2015, Nancy, France, August 23-26, 2015*. IEEE Computer Society, 2015, pp. 1156–1160. [Online]. Available: https://doi.org/10.1109/ICDAR.2015.7333942

[34] T. Q. Phan, P. Shivakumara, S. Tian, and C. L. Tan, "Recognizing text with perspective distortion in natural scenes," in *IEEE International Conference on Computer Vision, ICCV 2013, Sydney, Australia, December 1-8, 2013*. IEEE Computer Society, 2013, pp. 569–576. [Online]. Available: https://doi.org/10.1109/ICCV.2013.76

[35] A. Risnumawan, P. Shivakumara, C. S. Chan, and C. L. Tan, "A robust arbitrary text detection system for natural scene images," *Expert Syst. Appl.*, vol. 41, no. 18, pp. 8027–8048, 2014. [Online]. Available: https://doi.org/10.1016/j.eswa.2014.07.008

[36] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, "ASTER: an attentional scene text recognizer with flexible rectification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2035–2048, 2019. [Online]. Available: https://doi.org/10.1109/TPAMI.2018.2848939

[37] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, 1989. [Online]. Available: https://doi.org/10.1109/34.24792

[38] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[39] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006. International Society for Optics and Photonics, 2019, p. 1100612.