Context-based Contrastive Learning for Scene Text Recognition

Xinyun Zhang[1], **Binwu Zhu**[1], **Xufeng Yao**[1], **Qi Sun**[1], **Ruiyu Li**[2], **Bei Yu**[1]

[1]The Chinese University of Hong Kong [2]SmartMore

{xyzhang21,bwzhu,xfyao,qsun,byu}@cse.cuhk.edu.hk, royliruiyu@gmail.com

Abstract

Pursuing accurate and robust recognizers has been a long-lasting goal for scene text recognition (STR) researchers. Recently, attention-based methods have demonstrated their effectiveness and achieved impressive results on public benchmarks. The attention mechanism enables models to recognize scene text with severe visual distortions by leveraging contextual information. However, recent studies revealed that the implicit over-reliance of context leads to catastrophic out-of-vocabulary performance. On the contrary to the superior accuracy of the seen text, models are prone to misrecognize unseen text even with good image quality. We propose a novel framework, Context-based contrastive learning (ConCLR), to alleviate this issue. Our proposed method first generates characters with different contexts via simple image concatenation operations and then optimizes contrastive loss on their embeddings. By pulling together clusters of identical characters within various contexts and pushing apart clusters of different characters in embedding space, ConCLR suppresses the side-effect of overfitting to specific contexts and learns a more robust representation. Experiments show that ConCLR significantly improves out-of-vocabulary generalization and achieves state-of-the-art performance on public benchmarks together with attention-based recognizers.

Introduction

Reading text in the wild has been one of the most studied topics in the computer vision community. The rich information in scene text images plays a vital role in a series of artificial intelligence applications, such as Visual Question Answering (Biten et al. 2019), Autonomous Driving (Yu et al. 2021) and Image Retrieval (Gomez et al. 2018). Previous methods (Jaderberg et al. 2016; Wang, Babenko, and Belongie 2011) attempted to solve this problem from a symbol classification perspective. However, significant variations and even distortions in scene text images, such as blur and occlusion, hinder satisfactory performance. To bridge this gap, many attention-based methods (Fang et al. 2021; Yu et al. 2020; Yue et al. 2020; Li et al. 2019; Lyu et al. 2019; Qiao et al. 2020) have emerged and made remarkable progress in public benchmarks (Karatzas et al. 2015; Mishra, Alahari, and Jawahar 2012; Wang, Babenko, and Belongie 2011; Phan et al. 2013; Risnumawan et al. 2014; Karatzas et al. 2013). By leveraging the attention mechanism (Bahdanau, Cho, and Bengio 2015; Vaswani et al. 2017), models can attend to neighboring characters instead of looking at each one, leading to significant improvement in understanding irregular and hard-to-recognize text (Karatzas et al. 2015; Phan et al. 2013; Risnumawan et al. 2014).

SIMON

FDDB

Figure 1: Text recognition samples for in- and out-of-vocabulary text. For images containing in-vocabulary text (left), even confronted with occlusion, models can still predict correctly by inferring from the context, while models are more prone to make wrong predictions for images containing out-of-vocabulary text (right), even though the text is clear and free of distortions.

The key to the success of attention-based methods is to encode context information into character embeddings, whether in an auto-encoding or auto-regressive way. This feature allows models to reason target characters not only from the pixels at the corresponding position but also from the linguistic information coming from surrounding symbols. As shown in Figure 1 (left), even with severe occlusion that makes text non-identifiable, models can still infer the missing character based on the other ones. However, recent work (Wan et al. 2020) has revealed one crucial issue: Attention-based methods are more prone to vocabulary reliance. For images containing text seen in the training stage, state-of-the-art attention-based recognizers achieve promising accuracy while their performance drops drastically when predicting images with out-of-vocabulary text, even though they are relatively visually high-quality and free of distortions, as shown in Figure 1 (right). We conjecture that the leading cause is the over-reliance on context information. During training, the implicit context encoding dominates the discrimination process. Therefore, models overfit specific contexts instead of learning the discriminative features of each character. This problem significantly harms the robustness and generalization of scene text recognizers and heavily limits their application scenarios.

To mitigate this issue, our intuition is to learn a repre-

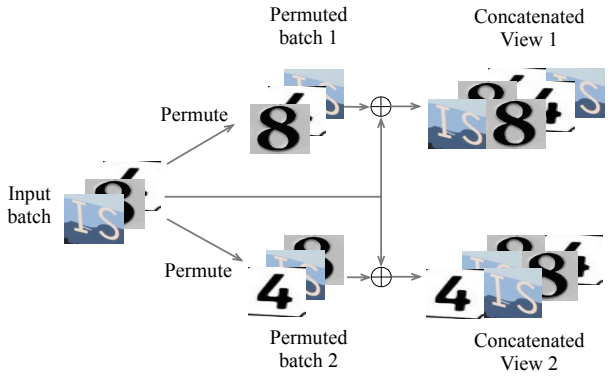


Figure 2: Context-based data augmentation. $\oplus$ denotes the concatenation operation. For characters in the input batch, their contexts are changed differently in two augmented views, for example, the context for 'I' in the first image of the original input is '-S', while in concatenated view 1 its context becomes '-S-8' and '4- -S' in view 2.

sentation that better balances the intrinsic character feature and context information to eliminate the over-reliance of the latter. Since the encoding process of the context is usually implicit, it is infeasible to manipulate the features directly. To this end, we propose context-based contrastive learning (ConCLR), a framework extending contrastive learning to the semantic space for scene text recognition (STR). In ConCLR, we contrast embeddings in different contexts to learn a representation more robust to context variations. To generate embeddings in various contexts, we propose context-based data augmentation (`ConAug`), a simple yet effective data augmentation technique to change on-image text's context. For most computer vision tasks, e.g., image classification, the context and foreground object are usually heterogeneous and tightly integrated, making it infeasible to manipulate the context without disturbing the foreground object. However, for STR, the context and foreground target to predict are both characters. Therefore, we can effortlessly modify the context for scene text images via simple image concatenation operations. As shown in Figure 2, given an input image, `ConAug` concatenates two different images to it to get two different views, respectively. Therefore, the character contexts in the original batch are differently changed in these two views. Then we feed these two views into the attention-based recognizers to get embeddings with augmented contexts. After being projected to another space, embeddings of identical characters are clustered while embeddings of different characters are pushed apart via optimizing the contrastive loss (Khosla et al. 2020). Through this, models are guided to learn character representations consistent in various semantic environments, which improves the generalization to unseen text.

Although out-of-vocabulary generalization (Wan et al. 2020) is essential for STR, it has been overlooked for years. The main reason is the commonly adopted training and evaluation settings. Conventionally, STR models are trained on two large synthetic datasets and evaluated on six real-world benchmarks. The vocabulary of the training set almost covers that of the evaluation set, as shown in Table 1. Therefore,

Table 1: In- and Out-of-vocabulary number of images for evaluation benchmarks. Training set MJ and ST comprises of 135272 words in its vocabulary. Please refer to the Experiment section for details of these datasets.

Benchmark	In-vocabulary number of images	Out-of-vocabulary number of images
IC13	1811	0
SVT	647	0
IIIT	2593	407
IC15	1487	324
CUTE	241	47
SVTP	645	0
OutText	0	1000

the performance is incapable of reflecting the generalization to unseen text. In light of this, we generate a new benchmark, OutText, consisting of 1000 images with pure random out-of-vocabulary text. We conduct experiments on both the commonly used benchmarks and OutText following conventional settings. Results demonstrate that, on the one hand, our method can significantly improve the generalization on unseen text; on the other hand, the performance on the seen text is also improved, suggesting the universal superiority of the learned representations. Further, we extend our method to combine with a language model (Fang et al. 2021) and achieves state-of-the-art performance on the public benchmarks.

To summarize, the major contributions of this paper are: First, we provide a new contrastive learning paradigm for STR, in which embeddings from different semantic contexts instead of visual augmentation are used for contrast. Second, based on this paradigm, we propose a framework, ConCLR, built on existing attention-based scene text recognizers to improve their generalization on unseen text. Third, we synthesize an out-of-vocabulary benchmark, OutText, to better reveal models' generalization to unseen text. Fourth, the extensive experiment results demonstrate the effectiveness of our proposed method. ConCLR significantly boosts the accuracy on unseen text and also achieves state-of-the-art performance on the public benchmarks, in which most text is seen in the training stage.

Related Work

Attention-based scene text recognition

Reading text in the wild is a challenging task due to the irregular layout and uncontrollable image quality. To this end, many attention-based methods (Shi et al. 2016; Yang et al. 2017; Cheng et al. 2017; Liu, Chen, and Wong 2018; Li et al. 2019; Qin et al. 2019; Baek et al. 2019; Wang et al. 2019; Yue et al. 2020; Wang et al. 2020; Yu et al. 2020; Fang et al. 2021) have been proposed and showed their superiority over previous CTC-based (Shi et al. 2016; Shi, Bai, and Yao 2017; Graves, Fernández, and Gomez 2006) and segmentation-based methods (Liao et al. 2020). As a feature alignment process, the attention mechanism can attend to relevant information for each character during the decoding stage, including the context. Previous models are mostly built on sequence-to-sequence architecture originated from NLP tasks, e.g., machine translation. (Shi et al. 2016) introduced a recurrent neural network (Bahdanau, Cho, and

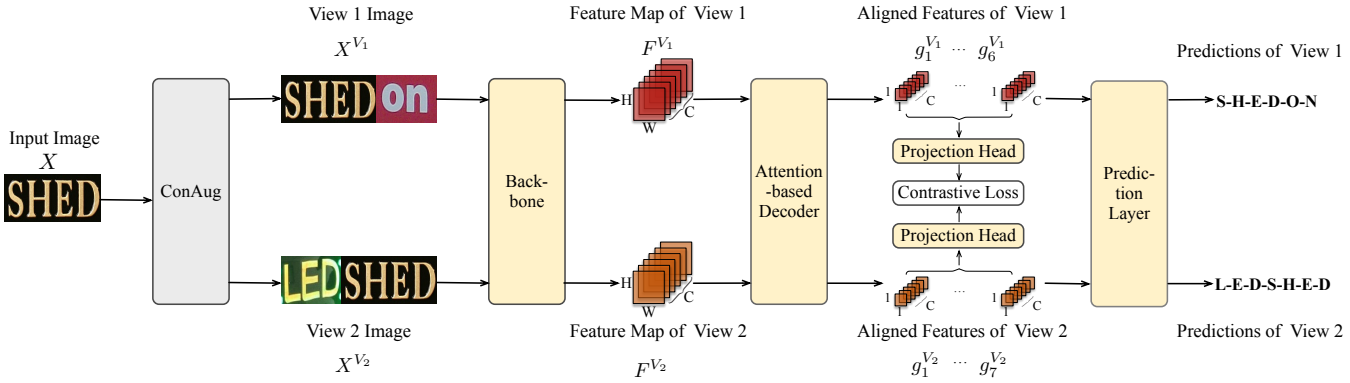


Figure 3: The main framework of ConCLR. Each input image is first fed into ConAug to get two context-based augmented views. Then these two views of an image are passed to the backbone and attention-based decoder to get aligned character features. We then pass these features to the projection head, and contrastive loss is optimized to pull together the positive samples and push apart the negative samples. Note that the forward process of the original batch is omitted for simplicity.

Bengio 2015) to align local pixels to an output sequence. Later, (Cheng et al. 2017) tailored a focus module to calibrate the attention location. Further, (Li et al. 2019) introduced a 2D attention mechanism for recognizing irregular text. (Yue et al. 2020) investigated the attention misalignment problem and designed a position enhancement module to fix it. However, because of the time-dependency of RNN-like structures, the inefficiency becomes a bottleneck for these auto-regressive methods. Thanks to the emergence of transformer (Vaswani et al. 2017), recent work (Fang et al. 2021; Yu et al. 2020) proposed parallel attention-based decoders, in which all characters are decoded simultaneously. Our framework is also built on this kind of recognizers.

Robust scene text recognition

The robustness of STR models, specifically in vocabulary, is a critical issue for applications. (Wan et al. 2020) firstly investigated vocabulary reliance and pointed out attention-based models suffer most from it. To remedy this, a mutual learning strategy (Wan et al. 2020) is proposed. They train a segmentation-based and attention-based model in parallel and align the features from the two models. By doing so, the segmentation-based model, generally with better out-of-vocabulary generalization, is used to calibrate the representation of the attention-based model. However, character-level annotations are required, which is an expensive cost for applications. In contrast, our method does not require any extra modules or character-level annotations to improve models' generalization on unseen text and performance on seen text.

Contrastive learning

Recent work (Chen et al. 2020; He et al. 2020; Grill et al. 2020) has significantly pushed the boundaries of representation learning by introducing contrastive learning. Generating positive samples via visual distortions and regarding other images as negative examples, (Chen et al. 2020; He et al. 2020) pull together embeddings of positive pairs and push apart that of negative pairs. Further, (Grill et al. 2020) proves merely using positive samples can also learn a promising embedding for downstream tasks. (Khosla et al. 2020) takes advantage of class labels as the criterion to separate positive and negative samples. For STR, (Aberdam et al. 2021) introduces a sub-word level contrastive learning framework, in which patches from different visually augmented images are considered as positive samples. Unlike all the methods mentioned above, instead of using visual augmentations to generate positive pairs, we propose to contrast characters within different semantic contexts in embedding space.

Method

The main framework of ConCLR is shown in Figure 3. Input images are first fed into ConAug to get context-based augmented views, then we pass these views into the network and get character embeddings. During training, on the one hand, these embeddings are used to transcript the text through the Prediction layer; on the other hand, we project them to a feature space where we conduct the contrastive loss. Now we detail the architecture and loss function in our framework.

Architecture

Context-based Data Augmentation. Our key insight is that by pulling together embeddings of the same character in different contexts and pushing apart embeddings of different characters, we can guide models to learn a representation better balances the intrinsic and context information. To create various contexts for given characters, we propose Context-based data augmentation (ConAug) in this section. Note that most in-the-wild text is horizontal; even for curved text, the reading order is still approximately from left to right. Drawing on this feature, ConAug leverages simple image concatenation operation to change text context. As shown in Figure 2, given a batch of images, ConAug first randomly permutes them twice, then concatenates the two permuted batches to the original one, respectively. We also discuss different ways of concatenation in the Experiment section. For each image, we get two different views after concatenation, in which the context is changed for characters therein.

Note that this operation requires no extra computation to manipulate the context, as a free lunch for STR. Besides,

the context modification bought by ConAug is additive. For each image, we add different characters to diversify the context while still preserving the original one. This feature still enables models to take advantage of the original context information.

Backbone. We adopt ResNet (He et al. 2016; Shi, Bai, and Yao 2017; Wang et al. 2020) as our backbone network. The output feature map size is 1/4 of the input image size. To capture long-range spatial dependencies, we also adopt the transformer unit (Vaswani et al. 2017).

Attention-based Decoder. Attention-based decoders align and aggregate relevant information and features for each character. The aligned embeddings are denoted as glimpses $\boldsymbol{g}$. During this process, the decoders are capable of involving context information to help infer target characters. According to the decoding pattern, we categorize the attention-based decoders into sequential decoders and parallel decoders, as shown in Figure 4.

Sequential decoders mostly adopt RNN-like structures and predict in an auto-regressive way, following the typical sequence-to-sequence framework. For time step t, the output is computed via the feature map $\boldsymbol{F}$ encoded by the backbone, the hidden state at the current time step $\boldsymbol{s}_t$, and the output $\hat{\boldsymbol{y}}_{t-1}$ from previous time step, which can be denoted as:

$$\boldsymbol{g}_t = f_{seq}(\boldsymbol{F}, \boldsymbol{s}_t, \hat{\boldsymbol{y}}_{t-1}), 0 < t \leq l, \tag{1}$$

where l is the sequence length. The prediction for each character depends on the previous one, and the decoding process continues until the output becomes the 'EOS' symbol. Despite the great success it has achieved, the time-dependent decoding strategy severely restricts its efficiency and makes the training process more tricky (Vaswani et al. 2017). Considering the fact that ConAug increases the length of the training data via image concatenation, the training efficiency problem will be significantly exaggerated if we use this type of decoders.

Inspired by Transformer (Vaswani et al. 2017), recent works (Yu et al. 2020; Fang et al. 2021) propose parallel attention-based decoders for STR. A fixed number of queries vectors, $\boldsymbol{q}$, are learned during training, each corresponding to the position encoding of a character order. Therefore, using the feature map $\boldsymbol{F}$ as the key, the glimpse vectors can be decoded in a parallel way, which can be denoted as:

$$(\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_l) = f_{par}(\boldsymbol{F}). \tag{2}$$

This parallel design remarkably improves decoders' efficiency in both training and evaluation stage. Besides, this architecture empowers models with more flexibility in attending to different spatial location, and shows its superiority especially on irregular benchmarks. In light of the advantages mentioned above, we use the parallel decoder proposed in (Fang et al. 2021) in our framework.

Projection Head. As mentioned in (Chen et al. 2020), directly contrasting the embeddings used to predict harms models' performance since we need to filter out irrelevant information in the features. Therefore, we use an auxiliary module denoted as $\texttt{proj}(\cdot)$ to map the representations to a space where contrastive loss is optimized. We conduct experiments on different architectures, e.g., identity mapping,

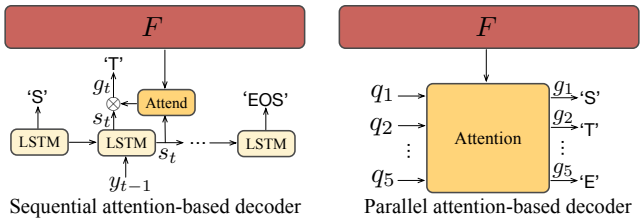


Figure 4: Attention-based decoders. The attention-based decoders can be mainly categorized into the sequential one (left) and the parallel one (right).

non-linear projection and linear projection, as shown in the Experiment section. Results show that linear projection is the best option.

Prediction Layer. We use a fully-connected (FC) layer to transcript the glimpse vectors to the probability of each character. Following the setting in previous work (Fang et al. 2021; Yu et al. 2020; Yue et al. 2020), our FC layer has 37 classes, including numbers 0-9, case-insensitive characters a-z, and one 'EOS' symbol. We denote the transcription process as:

$$(\hat{\boldsymbol{y}}_1, \hat{\boldsymbol{y}}_2, \cdots, \hat{\boldsymbol{y}}_l) = f_{pred}(\boldsymbol{g}_1, \boldsymbol{g}_2, \cdots, \boldsymbol{g}_l), \tag{3}$$

where $\hat{\boldsymbol{y}}$ is the predicted probability distribution for each character.

Loss function

There are two loss objectives in our framework, i.e., the recognition loss and the contrastive loss. The former, similar to (Fang et al. 2021), is used to train scene text recognizers while the latter is used to learn robust representations in semantic space. Before delving into them, we first clarify the notations. We define a batch of input data as $\{(\boldsymbol{X}_i, \boldsymbol{Y}_i), 0 < i \leq N\}$, where $\boldsymbol{X}_i$ is an input image, $\boldsymbol{Y}_i$ is the word-level label, and N is the batch size. Note that each $\boldsymbol{Y}_i$ can be further divided into character-level labels, denoted as $\boldsymbol{Y}_i = (\boldsymbol{y}_{i,1}, \boldsymbol{y}_{i,2}, \cdots, \boldsymbol{y}_{i,l_i})$, where l_i is the corresponding word length. After ConAug, the two augmented batches of data are denoted as $\{(\boldsymbol{X}_i^{V_1}, \boldsymbol{Y}_i^{V_1}), 0 < i \leq N\}$ and $\{(\boldsymbol{X}_i^{V_2}, \boldsymbol{Y}_i^{V_2}), 0 < i \leq N\}$, respectively.

For recognition loss, we compute it on the original input batch and the two augmented batches, which can be denoted as:

$$\begin{aligned}\mathcal{L}_{rec} = &\frac{1}{N}\sum_{i=1}^{N}(\sum_{o=1}^{l_i}\mathcal{L}_{ce}(\boldsymbol{y}_{i,o}, \hat{\boldsymbol{y}}_{i,o}) + \omega\sum_{p=1}^{l_i^{V_1}}\mathcal{L}_{ce}(\boldsymbol{y}_{i,p}^{V_1}, \hat{\boldsymbol{y}}_{i,p}^{V_1}) \\ &+ \omega\sum_{q=1}^{l_i^{V_2}}\mathcal{L}_{ce}(\boldsymbol{y}_{i,q}^{V_2}, \hat{\boldsymbol{y}}_{i,q}^{V_2})),\end{aligned} \tag{4}$$

where $\mathcal{L}_{ce}(\cdot)$ is cross-entropy loss, and ω is a hyper-parameter to tune the weight of the augmented samples for text recognition.

Contrastive loss is calculated on the two augmented batches. First, we pair up the two batches as: $\boldsymbol{I}^{aug} \equiv \{(\boldsymbol{X}_i^{V_1}, \boldsymbol{X}_i^{V_2}, \boldsymbol{Y}_i^{V_1}, \boldsymbol{Y}_i^{V_2}), 0 < i \leq N\}$. Given one pair of augmented data $\boldsymbol{T} \equiv (\boldsymbol{X}^{V_1}, \boldsymbol{X}^{V_2}, \boldsymbol{Y}^{V_1}, \boldsymbol{Y}^{V_2})$,

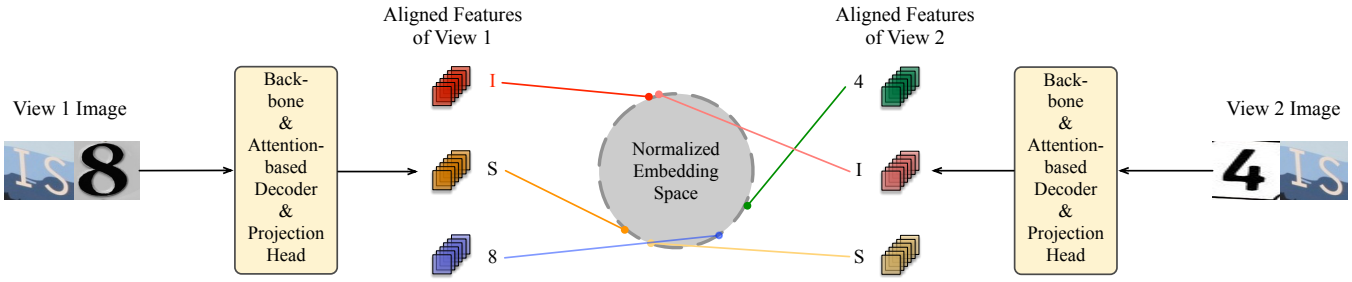


Figure 5: Context-based contrastive loss. For two augmented image views, we first extract the corresponding embeddings for each character via the backbone, attention-based decoder and the projection head. For each anchor, e.g., 'I' in red, we consider different 'I' among these embeddings as positive samples and other characters as negative samples.

the union of the aligned features of $\boldsymbol{X}^{V_1}$ and $\boldsymbol{X}^{V_2}$, can be denoted as $\mathcal{Z} \equiv \texttt{proj}(\boldsymbol{g}_1^{V_1}, \cdots, \boldsymbol{g}_{l^{V_1}}^{V_1}, \boldsymbol{g}_1^{V_2}, \cdots, \boldsymbol{g}_{l^{V_2}}^{V_2})$, and the union of the character labels can be denoted as $\mathcal{Y}^{aug} \equiv (\boldsymbol{y}_1^{V_1}, \cdots, \boldsymbol{y}_{l^{V_1}}^{V_1}, \boldsymbol{y}_1^{V_2}, \cdots, \boldsymbol{y}_{l^{V_2}}^{V_2})$. Let $m \in M \equiv \{1, \cdots, l^{V^1} + l^{V^2}\}$ be the index of any sample in $\mathcal{Z}$ or $\mathcal{Y}^{aug}$, $A(m) \equiv M \backslash \{m\}$ be the other indices except m itself, and $P(m) \equiv \{p \in A(m) : \boldsymbol{y}_p^{aug} = \boldsymbol{y}_m^{aug}\}$ be the indices of other aligned visual features having the same label as $\boldsymbol{z}_m$. Contrastive loss for one pair of data is defined as:

$$\mathcal{L}_{pair}(\boldsymbol{T}) = \sum_{m \in M} \frac{-1}{|P(m)|} \sum_{p \in P(m)} \log \frac{\exp(\boldsymbol{z}_m \cdot \boldsymbol{z}_p / \tau)}{\sum_{a \in A(m)} \exp(\boldsymbol{z}_m \cdot \boldsymbol{z}_a / \tau)}, \tag{5}$$

where the $\cdot$ symbol denotes the dot product, $\tau \in \mathbb{R}^+$ is a temperature hyper-parameter. Therefore, for a given batch, total contrastive loss can be calculated by:

$$\mathcal{L}_{clr} = \frac{1}{N} \sum_{\boldsymbol{T} \in \boldsymbol{I}^{aug}} \mathcal{L}_{pair}(\boldsymbol{T}). \tag{6}$$

The total loss takes the following form:

$$\mathcal{L}_{total} = \mathcal{L}_{rec} + \lambda \mathcal{L}_{clr}. \tag{7}$$

Here, λ, as the weight for contrastive loss, is a hyper-parameter. The calculation of the contrastive loss is shown in Figure 5. In our experiments, we set ω to 0.5, τ to 2, and λ to 0.2.

Experiments

In this section, we conduct extensive experiments to demonstrate the effectiveness of our proposed method. First, we detail the datasets used for training and evaluation. Then, the implementation details of our model are illustrated. Next, we show our results and compare them with the baseline parallel attention-based recognizers ABINet (Fang et al. 2021) on in-vocabulary and out-of-vocabulary text, respectively. For simplicity, we denote the parallel attention used in ABINet as ABINet-Vision. Then, we extend our framework with a language model used in (Fang et al. 2021), and achieve state-of-the-art performance. Finally, we conduct ablation studies to further investigate the functionality of each module.

Figure 6: Image samples of the seven benchmarks. For public benchmarks, IC13, IIIT, SVT are regular text benchmark, IC15, CUTE, and SVTP are irregular benchmarks. OutText is the synthesized benchmark of visually clear, out-of-vocabulary text.

Datasets and implementation details

To fairly compare our method with other state-of-the-art methods (Fang et al. 2021; Yu et al. 2020; Yue et al. 2020), we follow their settings for training and evaluation. The training set consists of two synthetic datasets, MJ (Jaderberg et al. 2016, 2014) and ST (Gupta, Vedaldi, and Zisserman 2016), and evaluation is conducted on six public benchmarks, including ICDAR 2013 (IC13) (Karatzas et al. 2013), ICDAR 2015 (IC15) (Karatzas et al. 2015), IIIT 5K-Words (IIIT) (Mishra, Alahari, and Jawahar 2012), Street View Text (SVT) (Wang, Babenko, and Belongie 2011), Street View Text-Perspective (SVTP) (Phan et al. 2013), and CUTE80 (CUTE) (Risnumawan et al. 2014), and our synthesized benchmark OutText.

OutText contains 1000 images. We paste random characters to the white background. Note that visual distortions, such as blur and occlusion, are excluded to guarantee the image quality. Considering that using `ConAug` lengthens the average word length of the training set, to exclude the impact of the word length, we synthesize OutText strictly following the word length distribution of MJ and ST. As shown in Figure 8, the word length concentrates between 3 to 8.

To reflect models' generalization to the unseen text, we count the word frequencies on training and evaluation set separately. As shown in Table 1, the training set MJ and ST consists of 135272 words as the vocabulary. For the evaluation sets, IC15, IIIT and CUTE consist of some out-of-vocabulary images while the words in IC13, SVT and SVTP are all covered in the training vocabulary. All the images in OutText contain text unseen in the training stage.

To ensure a fair comparison, we use the same experimental configuration as in ABINet (Fang et al. 2021). We use three transformer layers for the parallel attention module,

Table 2: Evaluation of the effectiveness of each module in ConCLR. For IC13-CUTE, the upper value represents the in-vocabulary accuracy while the lower one represents the out-of-vocabulary accuracy. In AVG, we calculate the overall average accuracy of the six public datasets. NA means there are no out-of-vocabulary images in the benchmark.

ConAug	ConLoss	IC13	SVT	IIIT	IC15	SVTP	CUTE	AVG	OutText
-	-	94.7 NA	90.1 NA	96.5 85.2	85.9 63.9	82.9 NA	88.4 76.6	89.8	63.2
✓	-	95.4 NA	89.9 NA	**96.8** 89.1	87.6 63.3	83.7 NA	**91.3** **87.2**	90.8	64.8
✓	✓	**95.9** NA	**92.1** NA	96.6 **89.7**	**88.7** **64.8**	**85.7** NA	90.0 85.1	**91.4**	**67.7**

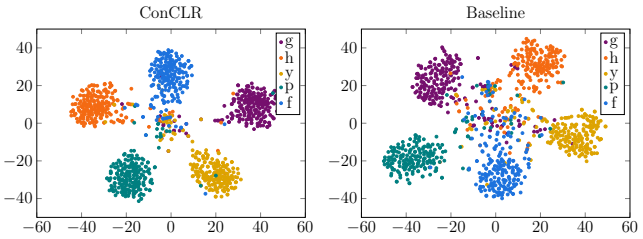


Figure 7: Embedding visualization of ConCLR and baseline parallel attention-based decoder. We randomly choose five characters from OutText, e.g., 'g', 'h', 'y', 'p' and 'f', and visualize all their corresponding embeddings using tSNE.

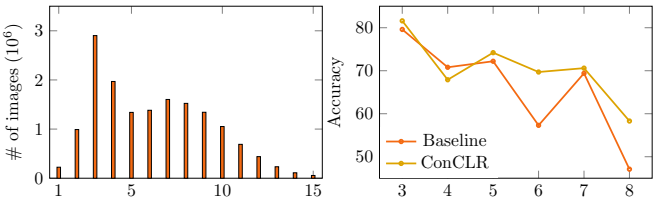
with eight heads for each of them. Images are resized to 32 $\times$ 128 with common data augmentation, such as random rotation, affine transformation, color jittering, and etc. We use ADAM as the optimizer, with a learning rate initialized to $1e^{-4}$ and decayed to $1e^{-5}$ at the 6-th epoch. All the experiments are conducted on four NVIDIA 2080Ti GPUs with batch size 384.

Analysis on seen and unseen data

We apply ConCLR on ABINet-Vision, and calculate the accuracy for seen and unseen data, respectively. The results are shown in the Table 2. For out-of-vocabulary data, ConCLR (using ConAug and contrastive loss both) significantly improves the performance. As we can see, compared with the vanilla parallel attention-based decoder, the accuracy on unseen data gains improvement of 4.5%, 8.5%, and 4.5% on IIIT, CUTE, and OutText, respectively. This indicates that ConCLR can guide to learn a representation that better balances the intrinsic information and the context information and is less dominated by the context. For IC15, the 0.9% improvement is relatively smaller, because of the poorer image quality as shown in Figure 6. When confronted with severely occluded or distorted images, it is infeasible for models to discriminate based on corresponding pixels for each character. In this case, the context information should be adopted to infer the target character. For in-vocabulary data, we can still observe notable performance improvement. ConCLR gains 1.2%, 2%, 0.1%, 2.8%, 2.8%, and 1.6% improvement on IC13, SVT, IIIT, IC15, SVTP and CUTE, respectively. This suggests the benefits brought by ConCLR are universal and not only restricted to the out-of-vocabulary data.

To further study the feature learned by ConCLR, we conduct embedding visualization, as shown in Figure 7. We sample embeddings of five characters in OutText and use

Figure 8: Left: Word length distribution of MJ and ST. The number of images with word length greater than 15 is very small, hence we do not show it in this figure. Right: The accuracy on words of different length. When word length is greater than eight, the advantage of ConCLR keeps increasing (omitted here). Please refer to the Appendix for the full data.



tSNE (Van der Maaten and Hinton 2008) to reduce their dimensions to two. As we can observe, the features learned from ConCLR are better clustered compared with features learned from baseline attention-based recognizers, demonstrating the superiority of the learned representation.

Analysis on word length

Since `ConAug` increases the average length of the training data, we also calculate the accuracy for different word lengths on OutText, as shown in Figure 8. For words with length greater than eight, ConCLR leads to overwhelming advantages benefiting from a wider word length distribution during training. For words with smaller lengths, ConCLR also has superiority on the average accuracy, suggesting that ConCLR guides models to learn a more representative embedding for words with different lengths, instead of overfitting to longer words.

Comparison with state-of-the-arts

To compare with previous arts, we also adopt a language model (LM) same as (Fang et al. 2021). Following the same experimental setting, we first use ConCLR to pretrain a vision model (ABINet-Vision), and then finetune with an LM. For a fair comparison, we reimplement ABINet and the results are shown in Table 3. As we can see, our method achieves state-of-the-art performance with a 0.8%, 0.4%, 0.5%, 0.8% and 3.8% improvement on IIIT, SVT, IC15, SVTP and CUTE, respectively. Especially for benchmarks containing out-of-vocabulary text, e.g., IIIT and CUTE, ConCLR shows its prominent superiority. For benchmarks containing no unseen text, ConCLR also achieves considerable improvement, demonstrating the benefit of the learned feature is universal.

Ablation study

Effectiveness of each module. The analysis of each module's effectiveness is shown in Table 2. By merely incorporating `ConAug` as a data augmentation technique, we can observe notable improvement on unseen data. Besides, this also outperforms the baseline on seen data. This suggests that we can reduce the overfitting to specific contexts and ameliorate the out-of-vocabulary generalization by simply diversifying the contexts. Further, the contrastive learning paradigm guides the model to learn a representation better balancing different features based on these various contexts,

Table 3: Results on IIIT5K, IC13, SVT, IC15, SVTP and CUTE datasets. † is our reimplementation.

Methods	Training Data	Annos	IIIT	IC13	SVT	IC15	SVTP	CUTE
ESIR (Zhan and Lu 2019)	MJ+ST	word	93.3	91.3	90.2	76.9	79.6	83.3
ASTER (Shi, Bai, and Yao 2017)	MJ+ST	word	93.4	91.8	89.5	76.1	78.5	79.5
RobustScanner (Yue et al. 2020)	MJ+ST	word	95.3	94.8	88.1	77.1	79.5	90.3
SAR (Li et al. 2019)	MJ+ST	word	91.5	91.0	84.5	69.2	76.4	83.3
DAN (Wang et al. 2020)	MJ+ST	word	94.3	93.9	89.2	74.5	80.0	84.4
SRN (Yu et al. 2020)	MJ+ST	word	94.8	95.5	91.5	82.7	85.1	87.8
SEED (Qiao et al. 2020)	MJ+ST	word	93.8	92.8	89.6	80.0	81.4	83.6
ABINet (Fang et al. 2021)	MJ+ST	word	96.2	97.4	93.5	86.0	89.3	89.2
ABINet-Vision†	MJ+ST	word	95.0	94.7	90.1	81.9	82.9	86.5
ABINet-Vision-ConCLR	MJ+ST	word	95.7	95.9	92.1	84.4	85.7	89.2
ABINet†	MJ+ST	word	95.7	**97.7**	93.9	84.9	88.5	87.5
ABINet-ConCLR	MJ+ST	word	**96.5**	**97.7**	**94.3**	**85.4**	**89.3**	**91.3**

Table 4: Ablation study on ConAug. For IC13-CUTE, the upper value represents the in-vocabulary accuracy while the lower one represents the out-of-vocabulary accuracy. In AVG, we calculate the overall average accuracy of the six public datasets. NA means there are no out-of-vocabulary images in the benchmark.

Concat	IC13	SVT	IIIT	IC15	SVTP	CUTE	AVG	OutText
SingleCat	95.4 NA	90.6 NA	96.8 88.4	87.6 64.2	83.3 NA	91.3 **87.2**	90.9	66.3
FixCat	95.2 NA	91.3 NA	**97.0** **89.9**	88.2 62.3	84.5 NA	**91.7** 85.1	91.2	67.2
RandCat	**95.9** NA	**92.1** NA	96.6 89.7	**88.7** **64.8**	**85.7** NA	90.0 85.1	**91.4**	**67.7**

which brings about larger improvement on both seen and unseen text.

Effectiveness of ConAug. Data augmentation plays an essential role in the contrastive learning framework. To explore its effectiveness, we design three concatenation patterns: *SingleCat*, for each input batch we only permute once and concatenate this permuted batch to the original batch on one random side, then the contrastive loss is computed on the concatenated batch and the original one; *FixCat*, for each input we permute twice and concatenate these two permuted batches to the original on one fixed side; *RandCat*, for each input batch we permute twice and concatenate these two batches to the original on one random side. The results are shown in Table 4. We can draw two conclusions: 1. Comparing *SingleCat* with *RandomCat*, we can observe 1.3%, 0.6% and 1.4% improvement of the unseen data on IIIT, IC15 and OutText, respectively. Concatenating with more images generates more diversified contexts and negative samples to contrast, which is beneficial for contrastive learning framework; 2. Comparing *FixCat* with *RandCat*, we can observe slight improvement on seen and unseen text. This indicates that position information is also context information, and we should not only change the concatenated characters but also their positions.

Effectiveness of the projection head. We also conduct ablation study on the architecture of the projection head. We consider three settings: 1. Identity mapping; 2. Non-linear projection head, we use a 512×256 FC layer, a 256×512 FC layer and a Relu activation in between; 3. Linear projection head, we use one fully-connected (FC) layer with dimension 512×512. As shown in Table 5, when using identity mapping as projection head, the contrastive loss slightly deteriorate the performance compared with the setting that we only use ConAug in Table 2, suggesting that directly contrast the embeddings from the backbone does not lead to beneficial representation. Besides, using non-linear or linear projection achieves remarkable improvement compared with identity mapping, which indicates that we should map the embedding to another space for contrasting. Further, by comparing the result of non-linear projection and that of the linear projection, we can observe that linear projection has some advantages, and we conjecture that non-linear projection may overly modify the original embeddings and weakens the effect of contrastive learning.

Table 5: Ablation study on the projection head. For IC13-CUTE, the upper value represents the in-vocabulary accuracy while the lower one represents the out-of-vocabulary accuracy. In AVG, we calculate the overall average accuracy of the six public datasets. NA means there are no out-of-vocabulary images in the benchmark.

Projection	IC13	SVT	IIIT	IC15	SVTP	CUTE	AVG	OutText
Identity mapping	95.4 NA	90.9 NA	96.7 **89.9**	86.9 61.7	83.9 NA	91.2 85.1	90.6	66.8
Non-linear projection	95.7 NA	90.3 NA	**97.0** 88.2	88.1 **65.4**	84.5 NA	**92.1** 85.1	91.2	67.4
Linear projection	**95.9** NA	**92.1** NA	96.6 89.7	**88.7** 64.8	**85.7** NA	90.0 85.1	**91.4**	**67.7**

Conclusion

In this paper, we propose a context-based contrastive learning framework for STR to improve attention-based scene text recognizers' generalization on unseen text. We use the ConAug module to create various contexts via simple image concatenation and then adopt contrastive loss on these character embeddings within different contexts. We pull together clusters of identical character embeddings in various contexts and push apart clusters of embeddings of different characters. Through this, models can learn a more discriminative representation that better balances the context and intrinsic information. Experiments show that our method can significantly improve out-of-vocabulary accuracy. Besides, our method also leads to remarkable improvement on the seen text and achieves state-of-the-art performance on six public benchmarks together with a language model.

References

Aberdam, A.; Litman, R.; Tsiper, S.; Anschel, O.; Slossberg, R.; Mazor, S.; Manmatha, R.; and Perona, P. 2021. Sequence-to-sequence contrastive learning for text recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Baek, J.; Kim, G.; Lee, J.; Park, S.; Han, D.; Yun, S.; Oh, S. J.; and Lee, H. 2019. What Is Wrong With Scene Text Recognition Model Comparisons? Dataset and Model Analysis. In *IEEE International Conference on Computer Vision (ICCV)*.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations (ICLR)*.

Biten, A. F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Valveny, E.; Jawahar, C.; and Karatzas, D. 2019. Scene text visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *International Conference on Machine Learning (ICML)*.

Cheng, Z.; Bai, F.; Xu, Y.; Zheng, G.; Pu, S.; and Zhou, S. 2017. Focusing Attention: Towards Accurate Text Recognition in Natural Images. In *IEEE International Conference on Computer Vision (ICCV)*.

Fang, S.; Xie, H.; Wang, Y.; Mao, Z.; and Zhang, Y. 2021. Read Like Humans: Autonomous, Bidirectional and Iterative Language Modeling for Scene Text Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gomez, L.; Mafla, A.; Rusinol, M.; and Karatzas, D. 2018. Single shot scene text retrieval. In *European Conference on Computer Vision (ECCV)*.

Graves, A.; Fernández, S.; and Gomez, F. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning (ICML)*.

Grill, J.-B.; Strub, F.; Altché, F.; Tallec, C.; Richemond, P.; Buchatskaya, E.; Doersch, C.; Avila Pires, B.; Guo, Z.; Gheshlaghi Azar, M.; Piot, B.; kavukcuoglu, k.; Munos, R.; and Valko, M. 2020. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Gupta, A.; Vedaldi, A.; and Zisserman, A. 2016. Synthetic data for text localisation in natural images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. In *Workshop on Deep Learning, Annual Conference on Neural Information Processing Systems (NIPS)*.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision (IJCV)*.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *13th International Conference on Document Analysis and Recognition (ICDAR)*.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *International Conference on Document Analysis and Recognition*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised Contrastive Learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Li, H.; Wang, P.; Shen, C.; and Zhang, G. 2019. Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition. In *AAAI Conference on Artificial Intelligence*.

Liao, M.; Pang, G.; Huang, J.; Hassner, T.; and Bai, X. 2020. Mask TextSpotter v3: Segmentation Proposal Network for Robust Scene Text Spotting. In *European Conference on Computer Vision (ECCV)*, 706–722.

Liu, W.; Chen, C.; and Wong, K.-Y. K. 2018. Char-net: A character-aware neural network for distorted scene text recognition. In *AAAI Conference on Artificial Intelligence*.

Lyu, P.; Yang, Z.; Leng, X.; Wu, X.; Li, R.; and Shen, X. 2019. 2D Attentional Irregular Scene Text Recognizer. arXiv:1906.05708.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Scene text recognition using higher order language priors. In *British Machine Vision Conference (BMVC)*.

Phan, T. Q.; Shivakumara, P.; Tian, S.; and Tan, C. L. 2013. Recognizing text with perspective distortion in natural scenes. In *IEEE International Conference on Computer Vision (ICCV)*.

Qiao, Z.; Zhou, Y.; Yang, D.; Zhou, Y.; and Wang, W. 2020. SEED: Semantics Enhanced Encoder-Decoder Framework for Scene Text Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qin, S.; Bissacco, A.; Raptis, M.; Fujii, Y.; and Xiao, Y. 2019. Towards unconstrained end-to-end text spotting. In *IEEE International Conference on Computer Vision (ICCV)*.

Risnumawan, A.; Shivakumara, P.; Chan, C. S.; and Tan, C. L. 2014. A robust arbitrary text detection system for natural scene images. *Expert Systems with Applications*.

Shi, B.; Bai, X.; and Yao, C. 2017. 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.*

Shi, B.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2016. Robust Scene Text Recognition With Automatic Rectification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Wan, Z.; Zhang, J.; Zhang, L.; Luo, J.; and Yao, C. 2020. On Vocabulary Reliance in Scene Text Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *IEEE International Conference on Computer Vision (ICCV)*.

Wang, P.; Yang, L.; Li, H.; Deng, Y.; Shen, C.; and Zhan, Y. 2019. A Simple and Robust Convolutional-Attention Network for Irregular Text Recognition. arXiv:1904.01375.

Wang, T.; Zhu, Y.; Jin, L.; Luo, C.; Chen, X.; Wu, Y.; Wang, Q.; and Cai, M. 2020. Decoupled attention network for text recognition. In *AAAI Conference on Artificial Intelligence*.

Yang, X.; He, D.; Zhou, Z.; Kifer, D.; and Giles, C. L. 2017. Learning to Read Irregular Text with Attention Mechanisms. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; and Ding, E. 2020. Towards accurate scene text recognition with semantic reasoning networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 12113–12122.

Yu, H.; Huang, Y.; Pi, L.; Zhang, C.; Li, X.; and Wang, L. 2021. End-to-end video text detection with online tracking. *Pattern Recognition*.

Yue, X.; Kuang, Z.; Lin, C.; Sun, H.; and Zhang, W. 2020. Robustscanner: Dynamically enhancing positional clues for robust text recognition. In *European Conference on Computer Vision (ECCV)*.

Zhan, F.; and Lu, S. 2019. ESIR: End-To-End Scene Text Recognition via Iterative Image Rectification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.