On Vocabulary Reliance in Scene Text Recognition

Zhaoyi Wan, Jielei Zhang, Liang Zhang, Jiebo Luo, Cong Yao
Megvii, China University of Mining and Technology, University of Rochester
i@wanzy.me, {yctmzjl,yaocong2010}@gmail.com, zhangliang04@hotmail.com, jluo@cs.rochester.edu

Abstract

The pursuit of high performance on public benchmarks has been the driving force for research in scene text recognition, and notable progress has been achieved. However, a close investigation reveals a startling fact that the state-of-the-art methods perform well on images with words within vocabulary but generalize poorly to images with words outside vocabulary. We call this phenomenon “vocabulary reliance”. In this paper, we establish an analytical framework to conduct an in-depth study on the problem of vocabulary reliance in scene text recognition. Key findings include: (1) Vocabulary reliance is ubiquitous, i.e., all existing algorithms more or less exhibit such characteristic; (2) Attention-based decoders prove weak in generalizing to words outside vocabulary and segmentation-based decoders perform well in utilizing visual features; (3) Context modeling is highly coupled with the prediction layers. These findings provide new insights and can benefit future research in scene text recognition. Furthermore, we propose a simple yet effective mutual learning strategy to allow models of two families (attention-based and segmentation-based) to learn collaboratively. This remedy alleviates the problem of vocabulary reliance and improves the overall scene text recognition performance.

1. Introduction

As a pivotal task in many visual recognition and comprehension systems, scene text recognition has been an active research field in computer vision for decades. Recently, the pursuit of high performance on benchmarks has drawn much attention from the community. Driven by deep learning and large volume of synthetic data, the recognition accuracy on standard benchmarks has escalated rapidly. For instance, the accuracy on IIIT-5k without lexicon has increased from 78.2% to 96.0% in a very short period.

Figure 1: The recurrent memory mechanism in RNN-attention based methods is actually a double-edged sword. The positive aspect is that for text images with words in the vocabulary (Left), even though image qualities are degraded (blur or partial occlusion), the content can be still correctly recognized. The negative aspect, which is previously neglected, lies in that for text images with words outside the vocabulary (Right), mistakes (marked in red) might easily occur.

However, an important issue has been overlooked for a long time: Even though achieving high accuracy on various benchmarks, state-of-the-art algorithms actually demonstrate obviously higher performance on images with words in the vocabulary than on those with words outside it. The gap is not caused by the image quality. As shown in Fig. 1, a top-performing text recognizer can correctly read the content even for images with poor quality but might make mistakes for images with better quality. The secret lies in the vocabulary: state-of-the-art methods seem inclined to memorize words that have been seen in the training phase. We call this phenomenon “vocabulary reliance”.

To further verify whether vocabulary reliance is common in scene text recognition, we reproduce a number of representative methods for scene text recognition, including CRNN, FAN, CA-FCN and ASTER. The same backbone network (ResNet-50) and training data (SynthText) are used for these methods, in order to rule out interference factors. As can be observed from Tab. 1, the performance gaps between test images with words in and outside the vocabulary are significant for all evaluated methods. It reveals that vocabulary reliance is ubiquitous.

In this paper, we systematically investigate the problem of vocabulary reliance in scene text recognition.
of vocabulary reliance in scene text recognition. An evaluation framework is established, in which training datasets with controlled vocabularies and targeted metrics are devised to assess and compare different module combinations.

Using training data with controlled vocabularies, we are able to inspect the impact of vocabulary on algorithm performance and abilities of different algorithms in learning language prior. Meanwhile, targeted metrics allow for the evaluation of the strengths and weaknesses of different module combinations in a quantitative and precise manner. Through experiments, we obtain a series of valuable observations and findings and accordingly give a few guidelines for choosing module combinations and suggestions for developing scene text recognition algorithms in the future.

Furthermore, in order to alleviate vocabulary reliance in existing methods, we propose a novel mutual learning strategy, which allows models with different PRED layers, i.e., attention-based decoder and segmentation-based decoder, to complement each other during training. Experimental results demonstrate its effectiveness in improving the accuracy and generalization ability of both attention decoders and segmentation-based methods.

The contributions of this work are as follows:
- We raise the problem of vocabulary reliance and propose an analytical framework for investigating it.
- We discovered through experiments the advantages and limitations of current PRED layers. Attention-based decoders generalize poorly from the learned vocabulary but perform well when trained on data with a random corpus. Segmentation-based methods can accurately extract visual features while the CTC-family generally has weaker visual observation ability.
- We found that the effect of CNTX modules, which perform context modeling, is highly coupled with PRED layers. We thus provide guidelines for choosing the CNTX modules according to PRED layers.
- Moreover, we present a simple yet effective mutual learning approach to allow models of different families to optimize collaboratively, which can alleviate the problem of vocabulary reliance.

## 2. Proposed Analytical Framework

In this section, we describe our analytical framework, including data, modules, and metrics, in detail.

<table>
<thead>
<tr>
<th>Methods</th>
<th>All</th>
<th>InVoc.</th>
<th>OutVoc.</th>
<th>Gap</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRNN [31]</td>
<td>86.8</td>
<td>91.1</td>
<td>68.7</td>
<td>22.5</td>
</tr>
<tr>
<td>FAN [2]</td>
<td>89.9</td>
<td>93.1</td>
<td>75.3</td>
<td>17.8</td>
</tr>
<tr>
<td>CA-FCN [23]</td>
<td>89.3</td>
<td>91.6</td>
<td>76.3</td>
<td>15.3</td>
</tr>
<tr>
<td>ASTER [33]</td>
<td>89.2</td>
<td>92.9</td>
<td>74.6</td>
<td>18.4</td>
</tr>
</tbody>
</table>

### 2.1. Test Data

To conduct experiments, we adopt various evaluation benchmarks, among which some are commonly used in prior works. We first briefly introduce public test datasets with real-word images, whose details are referred to [1].

**ICDAR2013 (IC13)** [15] is a dataset of ICDAR 2013 Robust Reading Competition for camera-captured scene text. **ICDAR2015 (IC15)** [14] comes from scene text images collected by Google glasses, where cropped text images are blurred, oriented and with low-resolution. **Street View Text (SVT)** [37] is an outdoor street images collection from Google Street View, including noisy, blurry or low-resolution images. **SVT Perspective (SVTP)** [28] focuses on curved text images. The dataset contains 645 evaluation images, which are severely distorted by non-frontal perspectives. **CUTE80 (CT)** [30] consists of 80 natural scene images, from which 288 cropped word images are generated for scene text recognition.

Basically, as shown in Fig. 1, the recognition of text images with difficulty in visual features, such as blur, stain, and irregular fonts, relies more on speculation according to the vocabulary. Thus, we group 5 datasets mentioned above into a set $\Omega$. The ground truths of $\Omega$ are collected as our corpus for synthetic training data. Therefore, $\Omega$ and its complement $\Omega'$ stand for the set of text images in and outside vocabulary, respectively.

Another evaluation dataset, **IIIT-5k (IIIT)** [27], is excluded from corpus collecting, which generally contains regular text and is of clear appearance. We choose IIIT as the stand-alone set to conduct the $\Omega'$ due to its relatively large amount of images and visual clearance. By the collected vocabulary, 1354 images in vocabulary are divided into $\Omega$ and the left 1646 images make $\Omega'$. They are named as IIIT-I and IIIT-O, respectively.

The size of the datasets and the number of their vocabularies are shown in Tab. 2. Besides, there are 3172 distinct words in the vocabulary of $\Omega$.

### 2.2. Training Data

Recent works for scene text recognition use synthetic data [7, 13] for training. **SynthText (ST)** is a dataset generated by a synthetic engine proposed in [7], whose background images are extracted from Google Image Search. It contains 80k images, from which researchers cropped about 7 million text instances for training.

As shown in Tab. 2, ST is generated from a large corpus from Newsroup20 [16] dataset, which has tens of thousands of words in the vocabulary. The large vocabulary of ST obfuscates the impact and cause of vocabulary reliance on such training data. Therefore we generate new training data for study by constraining the vocabulary.

Specifically, as stated in Sec. 2.1, our corpus is collected from test datasets. Using the synthetic engine of ST, three
datasets with a similar appearance and diverse corpus are conducted for thorough and controlled comparison. Examples are illustrated in Fig. 2.

**LexiconSynth (LS)** From collected ground truth words, we build the corpus for LS by uniformly sampling from instances. As the vocabulary of $\Omega$ is covered by LS, models trained with LS data acquire the facilitation of vocabulary learning when evaluated on $\Omega$. However, this purified corpus also exacerbates the over-fitting to words in vocabulary. In observation of the performance gap, properties about vocabulary learning of models can be dogged out.

**RandomSynth (RS)** In contrast to LS, the corpus of RS data is generated from characters in a random permutation. The lengths of the pseudowords are of the same distribution with those in LS, but the distribution of character classes is uniform. That is, the accuracy of models trained on RS is achieved without the assistance of vocabulary prior.

**MixedSynth (MS)** An intuitive solution for preventing algorithms from vocabulary reliance is to mix RS data into LS data. In our experiments, MS data is the union of LS and RS. Instances are sampled from RS and LS with ratio $r : (1 - r), r \in [0, 1]$. The training steps are fixed in all experiments. In comparison with datasets with a large vocabulary, the mixture of RS and LS is more practicable in real-world situations where the vocabulary is seldom completely given in advance.

**Synthesis Details** As the annotation of evaluation datasets serves in different manners on how to treat the case and punctuation of words, we collect the corpus as case-insensitive words without punctuation. During the rendering of LS data, each gathered word generates three instances with different variants: Uppercase, lowercase, and first-letter-capitalized case. Besides, words are inserted with a randomly chosen punctuation by a chance of 10%.

For the corpus of RS data, the proportion of letters, digits, and punctuation is about 6:3:1. Each word is rendered in the same three cases as LS data. Following the scale of ST, about 7 million cropped images are generated for RS and LS data respectively. When without special statements, the ratio $r$ of MS data is set as 0.5 empirically.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Voc.</th>
<th>Images</th>
<th></th>
<th></th>
<th></th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td>ST</td>
<td>7266715</td>
<td>-</td>
<td>76222</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>IC13</td>
<td>ST</td>
<td>857</td>
<td>158</td>
<td>549</td>
<td>142</td>
<td></td>
</tr>
<tr>
<td>IC15</td>
<td>ST</td>
<td>1369</td>
<td>442</td>
<td>669</td>
<td>348</td>
<td></td>
</tr>
<tr>
<td>SVT</td>
<td>ST</td>
<td>530</td>
<td>117</td>
<td>333</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>SVTP</td>
<td>ST</td>
<td>536</td>
<td>109</td>
<td>300</td>
<td>80</td>
<td></td>
</tr>
<tr>
<td>CT</td>
<td>ST</td>
<td>218</td>
<td>70</td>
<td>171</td>
<td>63</td>
<td></td>
</tr>
<tr>
<td>IIIT</td>
<td>ST</td>
<td>2429</td>
<td>571</td>
<td>1277</td>
<td>495</td>
<td></td>
</tr>
<tr>
<td>IIIT</td>
<td>Test</td>
<td>1354</td>
<td>1646</td>
<td>502</td>
<td>1270</td>
<td></td>
</tr>
</tbody>
</table>

Figure 2: Samples of generated training data. From top to bottom: all uppercase, all lowercase, and the first-letter-capitalized case. The left 2 columns are images picked up from LS, while the right 2 columns are ones from RS.

### 2.3. Module Combinations

According to [1], a typical scene text recognition method can be divided into four stages, transformation (TRAN), feature extraction (FEAT), context modeling (CNTX), and prediction (PRED). The CNTX stage is similar to sequence modeling (Seq.) in [1]. We extend to modeling context as we also take segmentation-based methods into consideration, for the sake of discussing the problem of vocabulary reliance in a broader perspective. The pipeline of scene text recognition is shown in Fig. 3.

In our experiments and analyses, we focus on CNTX and PRED stages, as these two stages are highly relevant to vocabulary reliance. TRAN and FEAT stages are fixed to control variables: No transformation layer is adopted and ResNet50 backbone is used in all combinations. Below, we will introduce three PRED layers and three choices for the CNTX stage.

**Prediction Layers** CTC [6] and attention-based decoders [3, 40] are two dominating approaches in the choices of prediction layers. As illustrated in Fig. 3d, CTC aligns the frame-wise predictions into the target string. Frames with the same characters without “BLANK”, which is introduced to stand for no characters, are removed in final outputs. CTC is widely used in many real-world applications [20] and academic researches [4, 9], due to its superior inference speed [1].

Attention-based (Atten. for short) decoders [2, 33] are state-of-the-art methods in the field of scene text recognition. A glimpse vector is generalized from the feature sequence, upon which an RNN is adopted to produce attention vectors over the feature sequence and produce character classification each in order (see Fig. 3c).

Recently, MaskTextSpotter [26] introduces instance segmentation to localize and classify each character separately and inspires following works [5, 23, 41]. Although segmentation-based (Seg. for short) methods directly extract characters by finding connected components in the segmentation map, the large receptive field of deep convolutional networks might bring vocabulary reliance.

**Context Modules** Bi-directional LSTM (BLSTM) [11] is employed for context modeling on top of feature maps ex-
As illustrated in Fig. 3e, the BLSTM module takes feature sequences as input, which are transformed from feature maps by pooling or convolution with strides. It is a common practice in many scene text recognition methods [34, 40] for context modeling, as the BLSTM scans and maps features in the bi-directional order.

Pyramid pooling module (PPM) [49] shown in Fig. 3f is another choice for context modeling, which is proved effective on segmentation-based methods [18]. It utilizes adaptive average pooling to pool feature maps into different square resolutions (1, 3, 4, 6 in our experiments). Pooled features are then resized to the input resolution by bi-linear interpolation and concatenated with original features to gain global context information in different scales.

Since segmentation-based methods are incompatible with BLSTM, PPM is a practical module for context modeling. Our experiments also validate its effectiveness in enhancing the vocabulary learning of models.

Besides, the explicit contextual modeling is not requisite for robust text recognition, as deep convolutional networks usually have large receptive fields [38, 47]. Though, in our experiments, context modeling modules do bring diversity in vocabulary learning and reliance.

The raw results are shown in Tab. 3, in which module combinations are named with circled numbers.

2.4. Metrics

Using our re-designed training data, we can evaluate the performance of algorithms on several training data. Several metrics are proposed for benchmarking the properties of models in aspects.

Firstly, we introduce a conventional metric for performance evaluation, General Accuracy (GA). The current practice for evaluating algorithms of scene text recognition is to evaluate models on public benchmarks with real-world images. We define the recognition accuracy on all test images of the mentioned evaluation datasets as GA, corresponding to the common evaluation in previous works.

In addition to the general metric, we further propose three specific metrics and their harmonic mean to fully reflect particular properties of different methods. For clarity, let’s define two functions. $Acc(X_{train}, X_{test})$ is the accuracy of models trained on dataset $X_{train}$ and tested on dataset $X_{test}$. $Gap(\cdot)$ is defined as the performance gap on IIIT-I and IIIT-O with the same training data $X_{train}$:

$$Gap(X_{train}) = Acc(X_{train}, IIIT-I) - Acc(X_{train}, IIIT-O).$$

Observation Ability (OA) Accurate visual feature extraction and recognition is the fundamental ability of scene text recognition methods. We define OA as how accurately an algorithm recognizes words without any vocabulary given
Table 3: The raw accuracy of models, which are numbered with circled numbers. “Gap” is the accuracy gap between IIIT-I and IIIT-O. “NGap” is normalized by recognition accuracy on IIIT.

<table>
<thead>
<tr>
<th>PRED</th>
<th>CNTX</th>
<th>No.</th>
<th>Data.</th>
<th>(\subseteq \Omega)</th>
<th>(\subseteq \Omega^c)</th>
<th>Gap/NGap</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RS</td>
<td>68.5</td>
<td>82.2</td>
<td>55.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MS</td>
<td>81.8</td>
<td>89.9</td>
<td>72.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
<td>85.7</td>
<td>92.7</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RS</td>
<td>70.3</td>
<td>84.6</td>
<td>57.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MS</td>
<td>81.6</td>
<td>88.6</td>
<td>71.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
<td>85.5</td>
<td>92.1</td>
<td>77.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>RS</td>
<td>68.6</td>
<td>82.4</td>
<td>55.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>MS</td>
<td>82.7</td>
<td>89.3</td>
<td>74.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>LS</td>
<td>87.0</td>
<td>92.7</td>
<td>79.8</td>
</tr>
</tbody>
</table>

Table 4: The computation of proposed metrics. Therein, \(Acc()\) and \(Gap()\) are defined in Sec. 2.4.

<table>
<thead>
<tr>
<th>Metrics</th>
<th>Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathcal{A})</td>
<td>(Acc(X_{train}, \Omega \cup \Omega^c))</td>
</tr>
<tr>
<td>(\Omega\mathcal{A})</td>
<td>(Acc(RS, \Omega \cup \Omega^c))</td>
</tr>
<tr>
<td>(\mathcal{V}\mathcal{A})</td>
<td>(Acc(LS, \Omega))</td>
</tr>
<tr>
<td>(\mathcal{V}\mathcal{G})</td>
<td>(1 - (Gap(LS) - Gap(RS)))</td>
</tr>
<tr>
<td>(\mathcal{S}\mathcal{M})</td>
<td>(3(x_\mathcal{A} + y_\mathcal{A} + z_\mathcal{A})^{-1})</td>
</tr>
</tbody>
</table>

in training data. In the context of our framework, \(\mathcal{OA}\) is measured by evaluating models trained on RS data with test images from all benchmarks (7406 images in total). As the recognition accuracy purely comes from the observation of visual features without learning any vocabulary, it indicates the ability of models to utilize visual observation.

Vocabulary Learning Ability (\(\mathcal{VA}\)) As stated in Sec. 1, it is likely for algorithms to employ learned vocabulary to refine or constrain recognition results of text images. Similar to \(\mathcal{OA}\), \(\mathcal{VA}\) is suggested for evaluating the recognition accuracy on limited vocabularies. In our experiments, measuring of \(\mathcal{VA}\) is to train models with LS data and evaluate the recognition accuracy on all images in \(\Omega\). \(\mathcal{VA}\) is meaningful for choosing models in text recognition tasks where lexicon is provided in advance.

Vocabulary Generalization (\(\mathcal{VG}\))

Human beings can easily generalize things from what they learnt, which inspires us to evaluate the vocabulary generalization (\(\mathcal{VG}\)) of an algorithm by measuring the performance of models trained with LS data on words out of vocabulary. In fact, we witness the vocabulary generalization of current recognition methods in our experiments. To fairly evaluate \(\mathcal{VG}\), the influence of image visual feature on the dataset, which brings an intrinsic gap between two image sets, is supposed to be eliminated. Therefore \(\mathcal{VG}\) is indicated by

\[
\mathcal{VG} = 1 - (Gap(LS) - Gap(RS)) \tag{2}
\]

where the score is subtracted from 1 in order to unify the monotonicity.

11429
Table 5: Metrics of models. The circled number corresponds to different combination of different module. No. is referred to Tab. 3.

<table>
<thead>
<tr>
<th>No.</th>
<th>PRED</th>
<th>(\overline{OA})</th>
<th>(VA)</th>
<th>(\overline{V\delta})</th>
<th>(OA)</th>
<th>(HM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>①</td>
<td>Atten.</td>
<td>81.0</td>
<td>85.7</td>
<td>77.1</td>
<td>69.6</td>
<td>76.9</td>
</tr>
<tr>
<td>②</td>
<td>Atten.</td>
<td>81.3</td>
<td>85.5</td>
<td>82.6</td>
<td>71.9</td>
<td>79.5</td>
</tr>
<tr>
<td>③</td>
<td>Atten.</td>
<td>83.1</td>
<td>87.0</td>
<td>79.1</td>
<td>69.8</td>
<td>82.0</td>
</tr>
<tr>
<td>④</td>
<td>CTC</td>
<td>75.8</td>
<td>77.8</td>
<td>92.4</td>
<td>65.8</td>
<td>77.1</td>
</tr>
<tr>
<td>⑤</td>
<td>CTC</td>
<td>80.1</td>
<td>84.8</td>
<td>89.5</td>
<td>63.5</td>
<td>77.5</td>
</tr>
<tr>
<td>⑥</td>
<td>CTC</td>
<td>78.4</td>
<td>79.9</td>
<td>90.2</td>
<td>67.6</td>
<td>78.1</td>
</tr>
<tr>
<td>⑦</td>
<td>Seg.</td>
<td>80.8</td>
<td>79.7</td>
<td>97.3</td>
<td>69.9</td>
<td>80.8</td>
</tr>
<tr>
<td>⑧</td>
<td>Seg.</td>
<td>81.3</td>
<td>81.6</td>
<td>94.0</td>
<td>70.5</td>
<td>80.9</td>
</tr>
</tbody>
</table>

Harmonic Mean (\(HM\)) For a overall metric, the harmonic mean of \(OA, VA, \overline{V\delta}\) is adopted as the summary score:

\[
HM = 3 \left( \frac{1}{\overline{OA}} + \frac{1}{VA} + \frac{1}{\overline{V\delta}} \right)^{-1}.
\]

\(HM\) can be taken as a standard for general comparison of different models.

Besides, evaluation on random string can be a metric, however, there is no standard benchmark that contains pure random labels with real-world complexity. Thus, it will not be discussed in this paper.

3. Comparisons and Analyses

Using our proposed framework in Sec. 2, we provide comparisons and analyses on various module combinations. Metrics of models are shown in Fig. 5. Based on the specific evaluation, we assess and analyze module combinations in different aspects.

3.1. Effect of Training Data

Fundamentally, we should first validate the effectiveness of the proposed dataset and explore the relevance of vocabulary reliance on training data. Experiments are conducted by gradually adjusting the ratio \(r\) in MS data from 0 to 1. Three models, ①, ④ and ⑦, are adopted for comparison. Besides the recognition accuracy on IIIT, we observe the probability of predicted words falling into the vocabulary, as shown in Fig. 4.

With RS data mixed into the LS data, recognition accuracy on IIIT is improved as models trained with the mixed data are less prone to be misled by vocabulary reliance. Especially for model ①, the recognition accuracy on IIIT increases from 77.8% to 84.4%, benefiting from the mixed RS data with a ratio of 25%.

The improvement in accuracy ceases when \(r\) reaches around 0.5. On one hand, the reduction of the probability to produce word prediction in vocabulary proves it effective to countervail vocabulary reliance with RS data. On the other hand, it requires a sufficient ratio of LS data to learn vocabulary from training data.

3.2. Comparison of Prediction Layers

From Fig. 5a, we perceive the consistent performance gap between models trained with RS, MS, and LS data, despite PRED layers nor CNTX modules. It shows that all the combinations suffer from the problem of vocabulary reliance, but the severity differs.

Moreover, we illustrate the performance gap on IIIT of model ①, ④ and ⑦ trained with different training data. The models are built without CNTX modules, using the Atten., CTC, and Seg. PRED layers, respectively. The attention-based decoder starts with the highest gap on the point where \(r = 0\) (LS data), as shown in Fig. 5b. With more RS data mixed into the training set, the gap of attention-based decoder decreases. The trend verifies the advantage of attention-based decoders on vocabulary learning and inferiority on vocabulary reliance.

In addition to vocabulary reliance, a thorough comparison of our proposed metrics of the PRED layers is illustrated in Fig. 6a. The performance of CTC is generally covered by the other two prediction layers, on metrics including both accuracy and generalization. Attention-based and segmentation-based decoders gain advantages in \(VA\) and \(\overline{V\delta}\) respectively. They also perform similarly well in \(\overline{OA}\), indicating the ability to accurate recognition according to visual features only.
3.3. Comparison of Context Modules

Basically, the adoption of context modules improves the vocabulary learning of models, as validated by the $\mathcal{V}A$ of module combinations. For example, PPM, which is not widely used in prior scene text recognition methods, brings boost on $\mathcal{VA}$ in combination with PRED layers: 3.9% for Seg. and 10.5% for CTC. On the other hand, as shown in Fig. 7, the strength in $\mathcal{VA}$ usually carries a decrease in $\mathcal{VG}$.

Similar to PRED layers, the evaluation results of CNTX modules are illustrated in Fig. 8a and Fig. 8b. We find that the effect of CNTX modules in detail is highly coupled with prediction layers.

As stated in Sec. 3.2, attention-based decoders are more powerful in learning vocabulary from training data. Consequently, it brings less change in $\mathcal{VA}$ and $\mathcal{VG}$ to add more context modules to attention-based PRED layers. Besides, context modules, which perform as contextual information extractor, in fact, facilitates visual observation of attention-based and segmentation-based decoders.

As for CTC-family models, the situation is different. PPM and BLSTM significantly improve their $\mathcal{VA}$ and impair the $\mathcal{VG}$, as the CTC decoder itself lacks of proper context modeling. The performance change in the three metrics brought by context modules on CTC-family models is shown in Fig. 8a.

In summary, it is effective to strengthen the vocabulary learning of models with proper context modules: BLSTM for attention-based, PPM for CTC and segmentation-based decoder. After all, it is a trade-off between $\mathcal{VA}$ and $\mathcal{VG}$.

3.4. Combination Recommendation

Based on Tab. 5 and the previous analyses, we recommend two combinations for different situations, depending on whether the vocabulary of target images are given.

Model ③, attention-based with BLSTM, achieves the best $\mathcal{VA}$ benefiting from the powerful CNTX module and PRED layer. This merit of model ③ in vocabulary learning also leads to the best $\mathcal{VG}$, corresponding to the performance on conventional benchmarks. It is evidenced by the high score in $\mathcal{VA}$ and $\mathcal{VG}$ that ③ can perform well in applications where the vocabulary of test images are mostly a restricted subset of training data. Accordingly, model ④, similar to [40] in network design, is our first recommended combination for strong vocabulary learning ability.

As for many applications in the industry, algorithms trained with data in limited vocabulary are supposed to generalize well to more general words. Model ⑦ maintains good vocabulary generalization ability as it gets the best $\mathcal{VG}$. Therefore, we recommend the combination ⑦, which is a CA-FCN-like [23] structure, for scenarios where the generalization of vocabulary is concerned.

4. Remedy by Mutual Learning

Previous sections demonstrate the trade-off between $\mathcal{VA}$ and $\mathcal{VG}$ and the diverse advantages of models. In this section, we propose a simple yet effective training strategy for combining advantages of models in different prediction layers, i.e., attention-based and segmentation-based decoders.

The idea is basically inspired by knowledge distillation [10] and deep mutual learning [48]. Similar to knowledge distillation, mutual learning of two models is a training strategy where models learn collaboratively. Knowledge distillation strategy transfers knowledge from a pre-trained powerful teacher network to student networks, while our approach optimizes two models simultaneously from scratch.

We choose the ensemble of the segmentation-based decoder and attention-based decoder as base models due to their advantages revealed in Fig. 6a. We suppose the generalization of segmentation-based decoders supervises attention-based decoders to learn to alleviate vocabulary reliance, and the accurate attention of attention-based decoders improves segmentation-based decoders in return.
4.1. Optimization

Let $\Theta_1$ and $\Theta_2$ be the network applying attention-based PRED layer and segmentation-based PRED layer, respectively. In addition to the original loss of the network $L_{\theta_1}$ and $L_{\theta_2}$, the Kullback Leibler (KL) Divergence is computed as an auxiliary loss. Then the ultimate loss for $\Theta_1$ and $\Theta_2$ are:

$$L_1 = \sum_y D_{KL}(p_1^y || p_1^y) + L_{\theta_1}$$
$$L_2 = \sum_y D_{KL}(p_2^y || p_2^y) + L_{\theta_2}$$

where $p_1$, $p_2$ are the sequence of logits produced by $\Theta_1$ and $\Theta_2$, respectively. $D_{KL}$ is the KL Divergence and $Y$ is the sequential label. Note that for segmentation-based decoders, the logits are “voted” scores [23] inside the shrunken region of characters.

From the Eq. 4, we can optimize the networks mutually supervised. The optimization is described in Alg. 1.

**Input**: Training data $X$ and label $Y$.

**Initialize**: $\Theta_1$ and $\Theta_2$ separately.

**while not converged do**

- $p_1 \leftarrow \text{Forward}(\Theta_1, X)$;
- $p_2 \leftarrow \text{Forward}(\Theta_2, X)$;
- Compute $L_1$ from Eq. 4;
- Backward($\Theta_1$, $L_1$);
- $p_1 \leftarrow \text{Forward}(\Theta_1, X)$;
- $p_2 \leftarrow \text{Forward}(\Theta_2, X)$;
- Compute $L_2$ from Eq. 4;
- Backward($\Theta_2$, $L_2$);

**end**

**Algorithm 1**: Optimization of mutual learning.

4.2. Experimental Validation

We evaluate the mutual learning strategy using the proposed evaluation framework and exhibit the raw accuracy and performance on our metrics in Tab. 3 and Tab. 6, respectively. Experimental results demonstrate the significant improvement of base models brought by the mutual learning strategy.

These two models united by the mutual learning strategy maintains diverse properties and distinguishable advantage. The joint training procedure combines their inclination to visual features and vocabularies by harmonizing their estimation with the KL divergence. As evidence indicates, the OA and VA of both models are improved, which verifies the effectiveness of the mutual learning strategy.

Moreover, the vocabulary reliance of attention-based decoder is neutralized by the segmentation-based decoder. In the training of attention-based decoder, the prediction of the segmentation-based model, which relies more on visual features, acts as an extra visual regularization. In addition to minimizing $L_{\theta_1}$, $\Theta_1$ is driven to fit the observation probability of $\Theta_2$. Quantitatively, the VA of $\Theta_1$ is boosted from 77.1% to 93.6%. In raw accuracy, the performance gap between images with words in and out of the vocabulary on LS data is almost halved (32.7% to 16.9%).

The qualitative comparison of the proposed mutual learning strategy is shown in Fig. 6b. Notable improvement on benchmarks demonstrates the effectiveness of the proposed mutual learning strategy, thus validating it reasonable to integrate the advantages of different PRED layers.

5. Conclusion

In this paper, we investigate an important but long-neglected problem: vocabulary reliance in scene text recognition methods. A comprehensive framework is built for comparing and analyzing individual text recognition modules and their combinations. Based on this framework, a series of key observations and findings have been acquired, as well as valuable recommendations, which could be conducive to the future research of scene text recognition. Moreover, we have analyzed current contextual and prediction modules and proposed a mutual learning strategy for enhancing their vocabulary learning ability or generalization ability to words out of vocabulary.

**Acknowledgement**

This research was supported by National Key RD Program of China (No. 2017YFA0700800). The authors would like to thank Minghui Liao for helpful discussions.
References


[4] Yunze Gao, Yingying Chen, Jinqiao Wang, Hanqing Lu, 1National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, and China. Reading scene text with attention convolutional sequence modeling. 1709. 3


[22] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. ArXiv, abs/1911.08947, 2019. 1

[23] Minghui Liao, Jian Zhang, Zhaoyi Wan, Fengming Xie, Jiajun Liang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Scene text recognition from two-dimensional perspective. In AAAI, 2019. 1, 2, 3, 7, 8


[33] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An and attentional scene and text recognizer and with flexible and rectification. In PAMI, pages 1–1, IEEE, 2018. 1, 2, 3, 4


