Decoupled Attention Network for Text Recognition

Tianwei Wang,[1] Yuanzhi Zhu,[1] Lianwen Jin,[1]* Canjie Luo,[1] Xiaoxue Chen,[1]
Yaqiang Wu,[2] Qianying Wang,[2] Mingxiang Cai[2]
[1]School of Electronic and Information Engineering, South China University of Technology
[2]Lenovo Research
{wangtw, z.yuanzhi, xxuechen}@foxmail.com, eelwjin@scut.edu.cn, canjie.luo@gmail.com,
{wuyqe, wangqya, caimx}@lenovo.com

Abstract

Text recognition has attracted considerable research interests because of its various applications. The cutting-edge text recognition methods are based on attention mechanisms. However, most of attention methods usually suffer from serious alignment problem due to its recurrency alignment operation, where the alignment relies on historical decoding results. To remedy this issue, we propose a decoupled attention network (DAN), which decouples the alignment operation from using historical decoding results. DAN is an effective, flexible and robust end-to-end text recognizer, which consists of three components: 1) a feature encoder that extracts visual features from the input image; 2) a convolutional alignment module that performs the alignment operation based on visual features from the encoder; and 3) a decoupled text decoder that makes final prediction by jointly using the feature map and attention maps. Experimental results show that DAN achieves state-of-the-art performance on multiple text recognition tasks, including offline handwritten text recognition and regular/irregular scene text recognition. Codes will be released.[1]

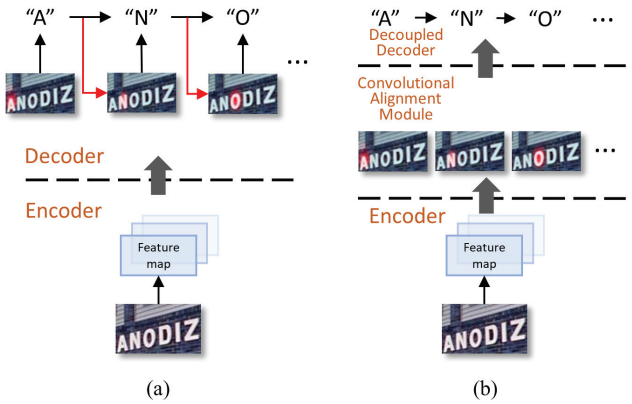

Figure 1: (a) Traditional attentional text recognizer, where the alignment operation is conducted using visual information and historical decoding information (red arrow). (b) Decoupled attention network, where the alignment operation is conducted using only visual information.

Introduction

Text recognition has drawn much research interest in recent years. Benefiting from the development of deep learning and sequence-to-sequence learning, many text recognition methods have achieved notable success (Long, He, and Yao 2018). Connectionist temporal classification (CTC) (Graves et al. 2006) and attention mechanism (Bahdanau, Cho, and Bengio 2015) are two most popular methods, among them attention mechanism shows significant better performance and has been studied frequently in recent years (Long, He, and Yao 2018).

The attention mechanism, proposed in (Bahdanau, Cho, and Bengio 2015) to tackle machine translation problem, was used to handle scene text recognition in (Lee and Osindero 2016; Shi et al. 2016), and since then it dominated text recognition with the following developments (Yang et al. 2017; Cheng et al. 2017; Bai et al. 2018; Luo, Jin, and Sun 2019; Li et al. 2019). The attention mechanism in text recognition is used to align and recognize characters, where the alignment operation has always been coupled with the decoding operation in previous work (Shi et al. 2016; Cheng et al. 2017; Bai et al. 2018; Li et al. 2019). As shown in Figure 1 (a), the alignment operation of traditional attention mechanism is carried out using two types of information. The first is a feature map that can be regarded as visual information from the encoder, and the second is historical decoding information (in the form of a recurrent hidden state (Bahdanau, Cho, and Bengio 2015; Luong, Pham, and Manning 2015) or the embedding vector of previous decoding result (Gehring et al. 2017; Vaswani et al. 2017)). The main idea underlying the attention mechanism is matching. Given a feature from the feature map, its attention score is computed by scoring how well it matches with the historical decoding information (Bahdanau, Cho, and Bengio 2015).

Traditional attention mechanism often encounters serious alignment problem (Cheng et al. 2017; Bai et al. 2018;

*Corresponding author

[1]https://github.com/Wang-Tianwei/Decoupled-attention-network

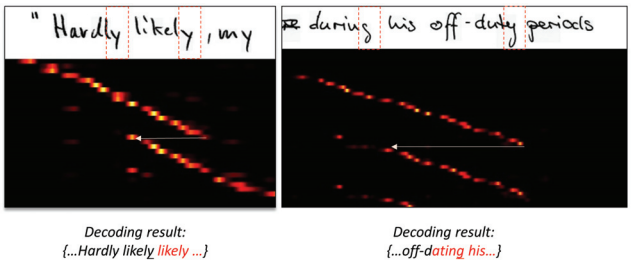

Figure 2: Visualization of fractional alignment of traditional attention mechanism (Bahdanau, Cho, and Bengio 2015; Shi et al. 2016) on long text.

Chorowski et al. 2015; Kim, Hori, and Watanabe 2017), This is because the coupling relationship inevitably leads to error accumulation and propagation. As shown in Figure 2, the matching-based alignment is easily affected by decoding result. In the left image, the two consecutive *"ly"* confuses matching operation; in the right image, the misrecognized result *"ing"* confuses matching operation. (Kim, Hori, and Watanabe 2017; Chorowski et al. 2015) also observed that attention mechanism struggles to align long sequence. Thus, it is intuitive to find a way to decouple the alignment operation from the historical decoding information, so that to reduce its negative impact.

To solve the aforementioned misalignment issue, in this paper we decouple the decoder of the traditional attention mechanism into an alignment module and a decoupled text decoder, and propose a new method called decoupled attention network (DAN) for text recognition. As shown in Figure 1 (b), compared with traditional attentional scene text recognizer, DAN needs no feedback from the decoding stage for alignment, thus avoiding the accumulation and propagation of decoding errors. The proposed DAN consists of three components including a feature encoder, a convolutional alignment module (CAM) and a decoupled text decoder. The feature encoder based on the convolutional neural network (CNN) extracts visual features from the input image. The CAM, substituting the traditional score-based recurrency alignment module, takes multi-scale visual features from the feature encoder as input, and generates attention maps with a fully convolutional network (Long, Shelhamer, and Darrell 2014) (FCN) in channel-wise manner. The decoupled text decoder makes the final prediction by using the feature map and attention maps with a gated recurrent unit (GRU) (Cho et al. 2014).

In summary, our contributions are summarized as follows:

- We propose a CAM to replace the recurrency alignment module in traditional attention decoders. The CAM conducts alignment operation from visual perspective, avoiding the use of historical decoding information, thus eliminating misalignment caused by decoding errors.
- We propose DAN, which is a effective, flexible (can be easily switched to adapt to different scenarios) and robust (more robust to text length variation and subtle disturbances) attentional text recognizer.
- DAN delivers state-of-the-art performance on several text recognition tasks, including handwritten text recognition and regular/irregular scene text recognition.

Related Work

Text recognition has attracted much research interest in the computer vision community. Early work of scene text recognition relied on low-level features, such as histogram of oriented gradients descriptors (Wang, Babenko, and Belongie 2011), connected components (Neumann and Matas 2012), etc. With the rapid development of deep learning, a large number of effective methods have been proposed. These methods can be mainly divided into two branches.

One branch is based on segmentation, it first detects characters then integrates characters into the output. (Bissacco et al. 2013) proposed a five hidden layers for character recognition and a n-gram approach for language modeling. (Wang et al. 2012) used a CNN to recognize characters and adopt a non-maximum suppression to obtain the final predictions. (Jaderberg, Vedaldi, and Zisserman 2014) proposed a weight-shared CNN for unconstrained text recognition. All of these methods require accurate individual detection of characters, which is very challenging.

The other branch is segmentation-free, it recognizes the text line as a whole and focuses on mapping the entire image directly to a word string. (Jaderberg et al. 2016) regraded scene text recognition as a 90k-class classification task. (Shi, Bai, and Yao 2017) modeled scene text recognition as a sequence problem by integrating the advantages of both deep convolutional neural network and recurrent neural network, and CTC was used to train the model end-to-end. (Lee and Osindero 2016) and (Shi et al. 2016) introduced attention mechanism to automatically align and translate words. From then on, more and more attention-based methods were proposed for text recognition. (Cheng et al. 2017) observed the attention drift problem and proposed a focusing net to draw back the drifted attention, but character-level annotation was required. (Bai et al. 2018) proposed a post-process, the edit probability to re-estimate the alignment; but they did not fundamentally solve misalignment. Focusing on recognition of irregular text, (Shi et al. 2016), (Luo, Jin, and Sun 2019) and (Zhan and Lu 2019) proposed to rectify text distortion and recognize the rectified text with an attention-based recognizer; (Liu, Chen, and Wong 2018) proposed to rectify text at the character level; (Yang et al. 2017) and (Liao et al. 2019) proposed to recognize text in two-dimensional perspective but character-level annotation is required; (Cheng et al. 2018) proposed to capture character feature in four directions. (Fang et al. 2018) proposed an attention and language ensemble network, and multiple losses from attention and language are accumulated for training it. (Li et al. 2019) proposed a simple and effective model using 2D attention mechanism.

Despite the notable success achieved by these attention-based methods, all of them consider attention to be a coupled operation between historical decoding information and visual information, and no study to date has focused on applying attention mechanism in long text recognition to the

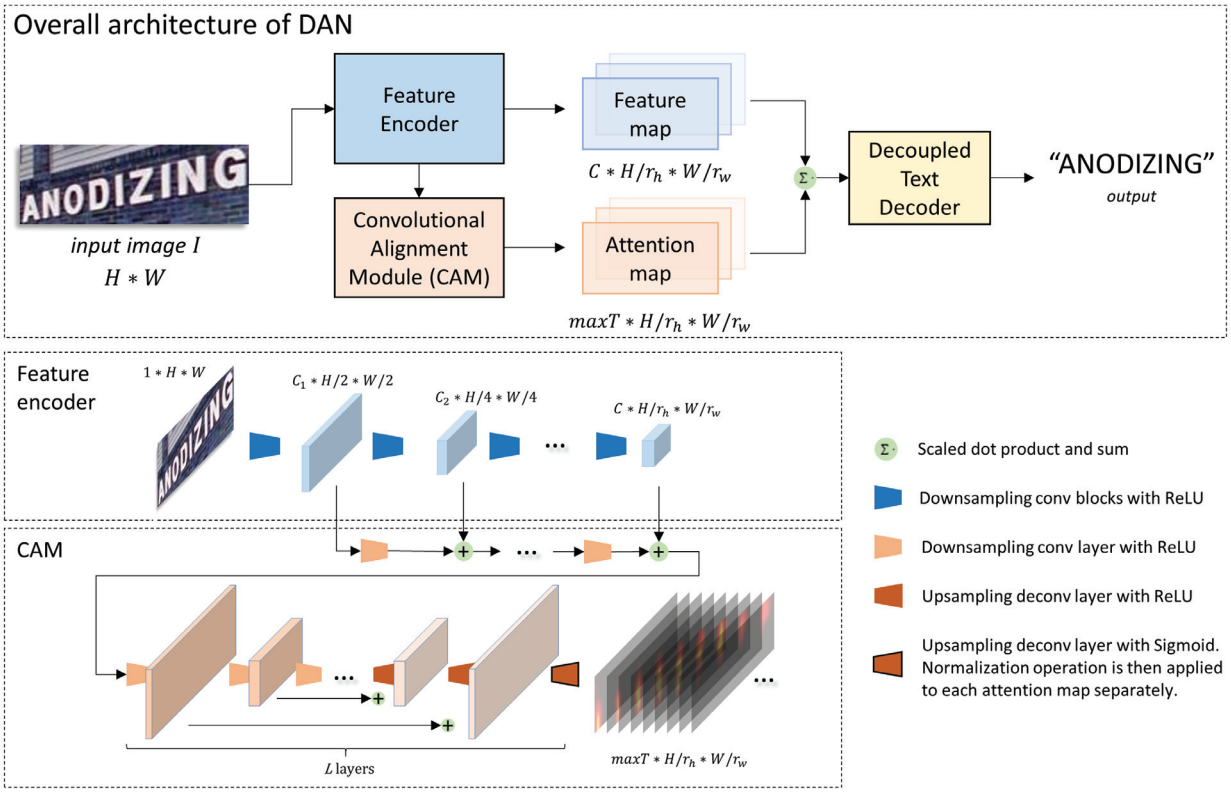

Figure 3: Overall architecture of DAN, and detailed architectures of the feature encoder and the CAM. The input image has a normalized height of H and a scaled width of W, C_1 and C_2 are the numbers of channels of the feature map.

best of our knowledge.

DAN

The proposed DAN aims at solving the misalignment issue of traditional attention mechanism through decoupling the alignment operation from using historical decoding results. To this end, we proposed a new convolutional alignment module (CAM) together with a decoupled text decoder to replace the traditional decoder. The overall architecture of DAN is illustrated in Figure 3. Details will be introduced in the followings.

Feature Encoder

We adopt a similar CNN-based feature encoder as previous study (Shi et al. 2018). The feature encoder $\mathcal{F}$ encodes the input image $\boldsymbol{x}$ of size $H \times W$ into feature map $\boldsymbol{F}$:

$$\boldsymbol{F} = \mathcal{F}(\boldsymbol{x}), \boldsymbol{F} \in \mathcal{R}^{C \times H/r_h \times W/r_w}. \quad (1)$$

where C , r_h and r_w denote the output channels, the height and the width downsampling ratio respectively.

Convolutional Alignment Module (CAM)

As shown in Figure 3, the input of our proposed CAM is visual features of each scale from the feature encoder. These multi-scale features are first encoded by cascade downsampling convolutional layers then summarized as input. Inspired by the FCN that makes dense predictions per-pixel channel-wise (*i.e.*, each channel denotes a heatmap of a class), we use a simple FCN architecture to conduct the attention operation channel-wise, which is quite different from current attention mechanism. The CAM has L layers; in the deconvolution stage, each output feature is added with the corresponding feature map from convolution stage. Sigmoid function with channel-wise normalization is finally adopted to generate attention maps $\boldsymbol{A} = \{\boldsymbol{\alpha_1}, \boldsymbol{\alpha_2}, ..., \boldsymbol{\alpha_{maxT}}\}$, where $maxT$ denotes the maximum number of channels, *i.e.*, the maximum number of decoding steps; and the size of each attention map is $H/r_h \times W/r_w$.

Compared with the FCN used for semantic segmentation, the CAM plays a completely different role to model a sequential problem. Although $maxT$ is pre-defined and should be fixed during training and testing, we will experimentally show that the setting of $maxT$ does not influence the final performance as long as it is reasonable.

By controlling the downsampling ratio r_h and change the stride of CAM, DAN can be flexibly switched between 1D and 2D form. When $H/r_h = 1$, DAN becomes a 1D recognizer and is suitable for long and regular text recognition; When $H/r_h > 1$(*e.g.*, for input image with height of 32, $r_h = 4$ results in a feature map with height of 4), DAN becomes a 2D recognizer and is suitable for irregular text recognition. Compared with previous 2D scene text recognizers, (Yang et al. 2017; Liao et al. 2019) which need character-level annotation for supervision; (Li et al. 2019)

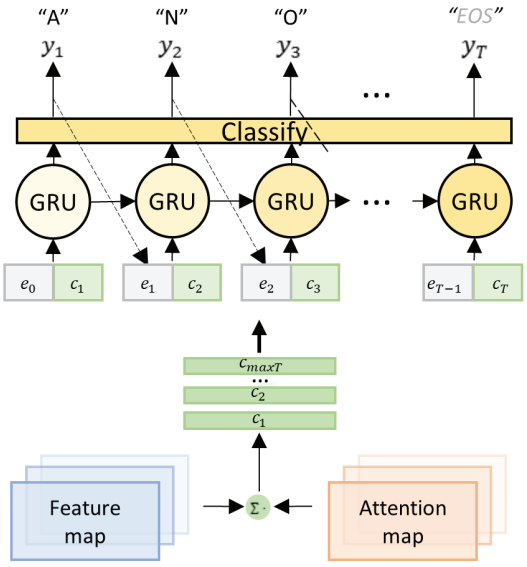

Figure 4: Detailed architecture of the decoupled text decoder. It consists of a GRU layer used to explore the contextual information and a linear layer to make predictions. '*EOS*' denotes end-of-sequence symbol.

which uses a tailored 2D attention for 2D spatial relationships caption, result in more complex than 1D form and has a poor performance on regular text recognition, DAN is significantly simple and flexible, while achieves state-of-the-art or comparable performance both in 1D (handwritten text) and 2D (irregular scene text) recognition.

Decoupled Text Decoder

Different from the traditional attentional decoder that conduct alignment and recognition concurrently, our decoupled text decoder takes encoded features and attention maps as input, and conducts recognition only. As shown in Figure 4, the decoupled text decoder computes context vector c_t as:

$$c_t = \sum_{x=1}^{W/r_w} \sum_{y=1}^{H/r_h} \alpha_{t,x,y} F_{x,y}. \tag{2}$$

At time step t, the classifier generates output y_t:

$$y_t = wh_t + b, \tag{3}$$

where h_t is the hidden state of the GRU, computed as:

$$h_t = GRU((e_{t-1}, c_t), h_{t-1}), \tag{4}$$

e_t is an embedding vector of the previous decoding result y_t. The loss function of DAN is as follows:

$$Loss = -\sum_{t=1}^{T} logP(g_t|I, \theta), \tag{5}$$

where θ and g_t denote all trainable parameters in the DAN and groudtruth at step t, respectively. Just like other attentional text recognizers, DAN uses word-level annotation for training.

Table 1: Detailed configuration of the feature encoder. 'Num' and 'hw' mean number of blocks and handwritten text recognition experiments, respectively.

Name	Configuration	Num	Downsampling Ratio		
			hw	scene-1D	scene-2D
Res-block0	3×3 conv	1	2×1	1×1	1×1
Res-block1	1×1 conv, 32 3×3 conv, 32	3	2×2	2×2	2×2
Res-block2	1×1 conv, 64 3×3 conv, 64	4	2×2	2×2	1×1
Res-block3	1×1 conv, 128 3×3 conv, 128	6	2×1	2×1	2×2
Res-block4	1×1 conv, 256 3×3 conv, 256	6	2×2	2×1	1×1
Res-block5	1×1 conv, 512 3×3 conv, 512	3	2×2	2×1	1×1

Performance Evaluation

In our experiments, two tasks are employed to evaluate the effectiveness of DAN, including handwritten text recognition and scene text recognition. The detailed network configuration of feature encoder is given in Table 1.

Offline Handwritten Text Recognition

Owing to its long sentences (up to 90 characters), diverse writing styles, and character-touching problem, the offline handwritten text recognition problem is highly complicated and challenging to solve. Therefore, it is a favorable testbed to evaluate the robustness and effectiveness of DAN.

For exhaustive comparison, we also conduct experiments on two popular attentional decoders: Bahdanau's attention (Bahdanau, Cho, and Bengio 2015) and Luong's attention (Luong, Pham, and Manning 2015). These attentional decoders are widely adopted for text recognition (Shi et al. 2018; Cheng et al. 2018; Luo, Jin, and Sun 2019; Li et al. 2019). When comparing with these decoders, the CAM and decoupled text decoder are replaced by them for the sake of fairness.

Datasets Two public handwritten datasets are used to evaluate the effectiveness of DAN, including IAM (Marti and Bunke 2002) and RIMES (Grosicki et al. 2009). The IAM dataset is based on handwritten English text copied from the LOB corpus. It contains 747 documents (6,482 lines) in the training set, 116 documents (976 lines) in the validation set and 336 documents (2,915 lines) in the test set. The RIMES dataset consists of handwritten letters in French. There are 1,500 paragraphs (11,333 lines) in the training set, and 100 paragraphs (778 lines) in the testing set.

Implementation Details On both databases we use the original whole-line training set with an open-source data-augmentation toolkit[2] to train the network. The height of the input image is normalized as 192 and the width is calculated with the original aspect ratio (up to 2048). To downsample the feature map into 1D, we add a convolution layer with kernel size 3×1 to the end of the feature encoder. $maxT$ is set to 150 in order to cover the longest line. The measure

[2]https://github.com/Canjie-Luo/Scene-Text-Image-Transformer

of performance is the Character or Word Error Rate (CER% or WER%), corresponding to the edit distance between the recognition result and groundtruth, normalized by the number of groundtruth characters (or words). At test time on RIMES dataset, we crop the test image with six pre-defined strategies (*e.g.*, {10,10} meant that the top 10 rows and the bottom 10 rows are cropped out), and then conduct recognition on them and the original image. A recognition score is calculated by averaging the output probabilities and the top scored one is chosen as the final result. All the layers of CAM except the last one are set as 128 channels in order to cover the longest text length. No language model or lexicon is used during experiments.

Table 2: Performance comparison on handwritten text datasets.

Methods	IAM		RIMES	
	WER	CER	WER	CER
(Salvador et al. 2011)	22.4	9.8	-	-
(Pham et al. 2014)	35.1	10.8	28.5	6.8
(Bluche 2016)	24.6	7.9	12.6	2.9
(Sueiras et al. 2018)	23.8	8.8	15.9	4.8
(Bhunia et al. 2019)[1]	**17.2**	8.4	10.5	6.4
(Zhang et al. 2019)	22.2	8.5	-	-
DAN	19.6	**6.4**	**8.9**	**2.7**

[1] Word-level recognition, where the words in the original image are cropped out then recognized.

Experimental Results As shown in Table 2, DAN exhibits superior performance on both datasets. On IAM dataset, DAN outperforms previous state-of-the-art by 1.5% on CER. Note that although (Bhunia et al. 2019) shows better performance on WER, their method needs cropped word images as input, while our method directly recognizes text lines. On RIMES, it is inferior to previous state-of-the-art by 0.2% on CER; but on WER, it has a great error reduction of 3.7% (relative error reduction of 29%). The great improvement in terms of WER indicates that DAN has a stronger capability of learning semantic information, which is helpful for long text recognition.

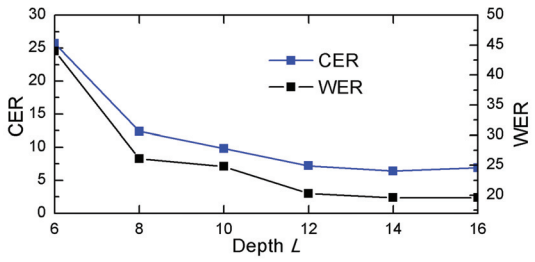

Figure 5: Performance comparison of different depth L on IAM dataset.

Ablation Study In this subsection, we will evaluate the influence of different depth L and output length $maxT$ of CAM.

Table 3: Performance comparison on different output lengths. The 'time/iter' means forward time per iteration on TITAN X GPU.

output length	IAM		time/iter
	WER	CER	
150	19.6	6.4	188.7 ms
200	19.5	6.3	189.5 ms
250	19.6	6.4	190.5 ms

Table 4: Performance comparison of different decoders. FE denotes the feature encoder of DAN. 'Bah' and 'Luong' denote Bahdanau's attention and Luong's attention, respectively.

Methods	IAM		RIMES	
	WER	CER	WER	CER
FE + Bah	25.9	9.9	9.1	3.0
FE + Luong	25.7	10.3	9.3	3.3
DAN	**19.6**	**6.4**	**8.9**	**2.7**

Output length: As shown in Table 3, different output lengths do not influence the performance, and the computation resource of additional channels is negligible, which indicates that DAN works well as long as the output length is reasonably set (longer than text length).

Depth: As shown in Figure 5, the performance of DAN degrades seriously as we reduce L, which show that the CAM should be deep enough to reach good performance. To successfully align one character, the reception field of CAM must be big enough to cover the corresponding features of this character and its neighbor regions.

Deep Insight into Eliminating Misalignments As shown in Table 4, compared with these two widely-used attentional decoders in the field of text recognition, DAN achieves significantly better performance.

To fine-grained study the improvements brought by the better alignment of DAN, we quantitatively discuss the relationship between obtained improvements of DAN and corresponding eliminated alignment errors. We propose a simple misalignment measurement method, which is based on the priori knowledge that all texts are written from left to right. This method consists of two steps: 1) picking the region with maximum attention score as attention center; 2) if current attention center is on the left side of the previous one, recording one misalignment. We divide the test samples into five groups by the text length: $[0, 30), [30, 40), [40, 50), [50, 60), [60, 70)$; each group contains more than 100 samples. In each group, the misalignments are added up then averaged to produce mean-misalignments per image (MM/img).

The experimental results are shown in Figure 6; The changes of CER improvement and eliminated misalignments are almost the same trend, which validates the performance gain of DAN relative to traditional attention comes from eliminating misalignments. In Figure 7, we show some visualization results of eliminated misalignments by our DAN.

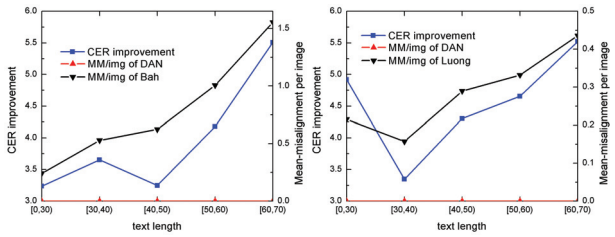

Figure 6: CER improvements of DAN on different text lengths and corresponding misalignments. 'Bah' and 'Luong' denote Bahdanau's attention and Luong's attention, respectively.

Error Analysis Figure 8 shows some typical error samples of DAN. In Figure 8 (a), the character 'e' is recognized as 'p' because of its confusing writing style. The misclassified 'p' is challenging for humans without contextual information. In Figure 8 (b), a space symbol is missed by the recognizer, because the two relevant words are too close. In Figure 8 (c), some noise texture is recognized as a word by DAN. However, DAN is still more robust than traditional attention on these samples. In Figure 8 (c) the confusing noises disturb the alignment operation of traditional attention and lead to unpredictable errors, while DAN is robust in alignment even if extra results are generated. Considering that the noises have almost the same texture with normal text, this type of error is very difficult to avoid, especially for DAN which conduct alignment only based on visual features.

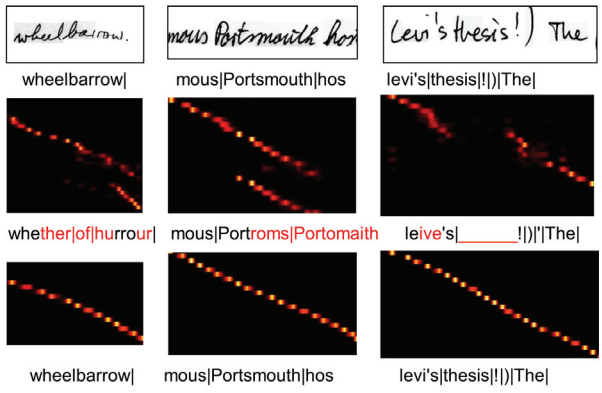

Figure 7: Visualization of attention maps and recognition results on IAM dataset. Top: original fractional images and corresponding groundtruth; middle: attention maps and recognition results of traditional attention; bottom: attention maps and recognition results of DAN.

Scene Text Recognition

Scene text recognition often encounters problems owing to the large variations in the background, appearance, resolution, text font, and so on. In this section, we will study the effectiveness and robustness of DAN on seven datasets including regular scene text datasets and irregular scene text

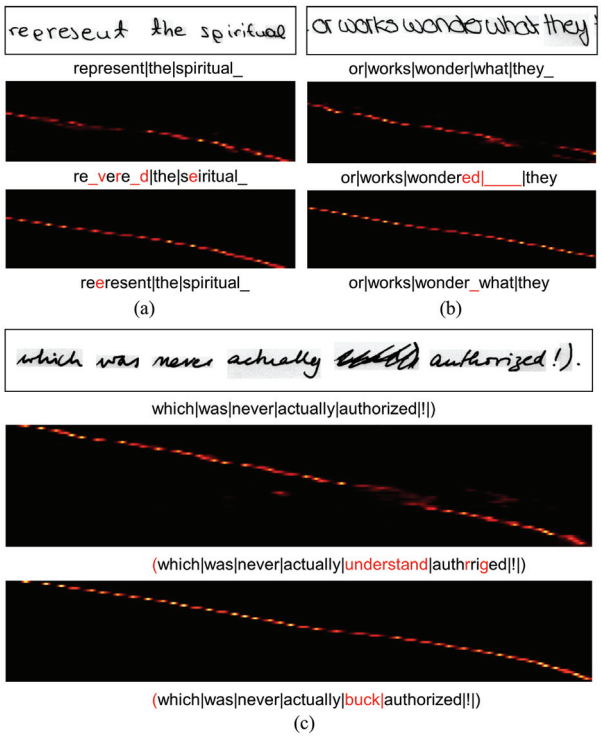

Figure 8: Visualization of typical error samples of DAN. The order of images is same as Figure 7. (a) Substitute error where character 'p' is misrecognized as 'e'; (b) delete error where a space symbol is missed; (c) insert error where some textures are recognized as 'buck'.

datasets. We will validate the performance of DAN in 1D and 2D form (denote as DAN-1D and DAN-2D); the detailed configurations of feature encoder are shown in Table 1.

Datasets Two types of datasets are used for scene text recognition: regular scene text datasets, including IIIT5K-Words (Mishra, Alahari, and Jawahar 2012), Street View Text (Wang, Babenko, and Belongie 2011), ICDAR 2003 (Lucas et al. 2003) and ICDAR 2013 (Karatzas et al. 2013); and irregular scene text datasets, including SVT-Perspective (Neumann and Matas 2012), CUTE80 (Risnumawan et al. 2014) and ICDAR 2015 (Karatzas et al. 2015).

IIIT5k was collected from the Internet, and contained 3,000 cropped word images for testing.

Street View Text (SVT) was collected from the Google Street View, and contained 647 word images for testing.

ICDAR 2003 (IC03) contained 251 scene images that are labeled with text bounding boxes. The dataset contained 867 cropped images.

ICDAR 2013 (IC13) inherited most images from IC03 and extends it with some new images. It consisted of 1,015 cropped images without associated lexicon.

SVT-Perspective (SVT-P) was collected from the side-view angle snapshots in Google Street View, and contained 639 cropped images for testing .

Table 5: Performance comparison on regular and irregular scene text datasets. 'Rect' represents rectification-based methods; '2D' represents 2D-based methods.

Methods	Rect	2D	Regular				Irregular		
			IIIT5k	SVT	IC03	IC13	SVT-P	CUTE80	IC15
(Cheng et al. 2017) [1]			87.4	85.9	94.2	93.3	-	-	70.6
(Cheng et al. 2018)			87.0	82.8	91.5	-	73.0	76.8	68.2
(Bai et al. 2018) [1]			88.3	87.5	94.6	**94.4**	-	-	73.9
(Liu et al. 2018)			89.4	87.1	94.7	94.0	73.9	62.5	-
(Shi et al. 2018)	✓		93.4	89.5	94.5	91.8	78.5	79.5	76.1
(Fang et al. 2018)			86.7	86.7	94.8	93.5	-	-	71.2
(Luo, Jin, and Sun 2019)	✓		91.2	88.3	95.0	92.4	76.1	77.4	68.8
(Liao et al. 2019) [1]		✓	92.0	86.4	-	91.5 [1]	-	79.9	-
(Li et al. 2019)		✓	91.5	84.5	-	91.0	76.4	83.3	69.2
(Xie et al. 2019)		✓	-	-	-	-	70.1	82.6	68.9
(Zhan and Lu 2019)	✓		93.3	**90.2**	-	91.3	79.6	83.3	**76.9**
DAN-1D			93.3	88.4	**95.2**	94.2	76.8	80.6	71.8
DAN-2D		✓	**94.3**	89.2	95.0	93.9	**80.0**	**84.4**	74.5

[1] character-level annotation required.

Table 6: Robustness study. 'ac': accuracy; 'gap': the gap between the original dataset; 'ratio': accuracy decreasing ratio.

Methods	IIIT	IIIT-p			IIIT-r-p			IC13	IC13-ex			IC13-r-ex		
	ac	ac	gap	ratio	ac	gap	ratio	ac	ac	gap	ratio	ac	gap	ratio
CA-FCN	92.0	89.3	-2.7	2.9%	87.6	**-4.4**	**4.8%**	91.4	87.2	-3.7	4.1%	83.8	**-6.9**	7.6%
DAN-1D	93.3	91.5	**-1.8**	**1.9%**	88.2	-5.1	5.4%	**94.2**	**91.2**	**-3.0**	**3.2%**	**86.9**	-7.3	7.7%
DAN-2D	**94.3**	**92.1**	-2.2	2.3%	**89.1**	-5.2	5.5%	93.9	90.4	-3.5	3.7%	**86.9**	-7.0	**7.5%**

CUTE80 focused on curved text, and consisted of 80 high-resolution images taken in natural scenes. This dataset contained 288 cropped natural images for testing.

ICDAR 2015 (IC15) contained 2,077 cropped images. A large proportion of images were blurred and multi-oriented.

Implementation Details We train our model on synthetic samples released by (Jaderberg et al. 2014) and (Gupta, Vedaldi, and Zisserman 2016). For better comparison, we compare DAN only with the methods that had also used these two synthetic datasets. The height of the input image is set to 32 and the width is calculated with the original aspect ratio (up to 128). $maxT$ is set as 25; L is set as 8; and all the layers of CAM except the last one are set as 64. We use the bi-directional decoder proposed in (Shi et al. 2018) for final prediction. channels. With ADADELTA (Zeiler 2012) optimization method, the learning rate is set as 1.0 and reduced to 0.1 after the third epoch.

Experimental Results As shown in Table 5, DAN achieves state-of-the-art or comparable performance on most datasets. For regular scene text recognition, DAN achieves state-of-the-art performance on IIIT5K and IC03, and is just a little behind the current state-of-the-art on SVT and IC13. DAN-1D performs a little better on IC03 and IC13, because images from these two datasets are usually clean and regular. For irregular scene text recognition, the most advanced methods can be divided into two types: rectification based and 2D based. DAN-2D achieves state-of-the-art performance on SVT-P and CUTE80, and it exhibits the best performance among 2D recognizers.

Robustness Study Scene text is usually affected by environmental disturbances. To check whether DAN is sensitive to subtle disturbances, we also conduct robustness study on IIIT-5k and IC13 datasets, and compare DAN with the most-recent 2D scene text recognizer, CA-FCN (Liao et al. 2019). We add some disturbances on these two datasets as follows:

IIIT-p: Padding the images in IIIT5k with extra 10% height vertically and 10% width horizontally by repeating the border pixels. **IIIT-r-p**: 1. Separately stretching the four vertexes of the images in IIIT5k with a random scale up to 20% of height and width respectively. 2. Repeating border pixels to fill the quadrilateral images. 3. Transforming the images back to axis-aligned rectangles. **IC13-ex**: Expanding the bounding boxes of the images in IC13 to expanded rectangles with extra 10% height and width before cropping. **IC13-r-ex**: 1. Expanding the bounding boxes of the images in IC13 randomly with a maximum 20% of width and height to form expanded quadrilaterals. 2. The pixels in axis-aligned circumscribed rectangles of those images are cropped.

The results are shown in Table 6. In most cases DAN exhibits to be more robust than CA-FCN, which again validates its robustness.

Discussion

Advances of DAN: 1) **Simple**. DAN uses off-the-shelf components; all of them are easy to implement. 2) **Effective**. DAN achieves state-of-the-art performance on multiple text

recognition tasks. 3) **Flexible**. The form of DAN can be easily switched between 1D and 2D. 4) **Robust**. DAN exhibits more reliable alignment performance when facing long text. It is also more robust facing subtle disturbances.

Limitations of DAN: The CAM uses only visual information for alignment operation; thus when it comes text-like noises, it struggles to align the text. This kind of error is shown in Figure 8 (c) and may be a common issue for most attention mechanism.

Conclusion

In this paper, an effective, flexible and robust decoupled attention network is proposed for text recognition. To address the misalignment issue, DAN decouples the decoder of the traditional attention mechanism into a convolutional alignment module and a decoupled text decoder. Compared with the traditional attention mechanism, DAN effectively eliminates the alignment errors and achieves the state-of-the-art performance. Experimental results on multiple text recognition tasks have shown its effectiveness and merit. Particularly, DAN shows significant superiority when dealing with long text recognition, such as handwritten text recognition.

Acknowledgement

This research is supported in part by NSFC (Grant No.: 61936003), the National Key Research and Development Program of China (No. 2016YFB1001405), GD-NSF (no.2017A030312006), Guangdong Intellectual Property Office Project (2018-10-1), and GZSTP (no. 201704020134).

References

Bahdanau, D.; Cho, K.; and Bengio, Y. 2015. Neural machine translation by jointly learning to align and translate. In *ICLR*.

Bai, F.; Cheng, Z.; Niu, Y.; Pu, S.; and Zhou, S. 2018. Edit probability for scene text recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1508–1516.

Bhunia, A. K.; Das, A.; Bhunia, A. K.; Kishore, P. S. R.; and Roy, P. P. 2019. Handwriting recognition in low-resource scripts using adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 4767–4776.

Bissacco, A.; Cummins, M.; Netzer, Y.; and Neven, H. 2013. Photoocr: Reading text in uncontrolled conditions. In *IEEE International Conference on Computer Vision*, 785–792.

Bluche, T. 2016. Joint line segmentation and transcription for end-to-end handwritten paragraph recognition. In *Annual Conference on Neural Information Processing Systems*, 838–846.

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*, 5086–5094.

Cheng, Z.; Xu, Y.; Bai, F.; Niu, Y.; Pu, S.; and Zhou, S. 2018. AON: Towards arbitrarily-oriented text recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5571–5579.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 1724–1734.

Chorowski, J.; Bahdanau, D.; Serdyuk, D.; Cho, K.; and Bengio, Y. 2015. Attention-based models for speech recognition. In *Annual Conference on Neural Information Processing Systems*, 577–585.

Fang, S.; Xie, H.; Zha, Z.-J.; Sun, N.; Tan, J.; and Zhang, Y. 2018. Attention and language ensemble for scene text recognition with convolutional sequence modeling. In *ACM Multimedia*, 248–256.

Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, 1243–1252.

Graves, A.; Fernández, S.; Gomez, F. J.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *International Conference on Machine Learning*, 369–376.

Grosicki, E.; Carré, M.; Brodin, J. M.; and Geoffrois, E. 2009. Results of the rimes evaluation campaign for handwritten mail processing. In *IAPR International Conference on Document Analysis and Recognition*, 941–945.

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*, 2315–2324.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Synthetic data and artificial neural networks for natural scene text recognition. In *Annual Conference on Neural Information Processing Systems Deep Learning Workshop*.

Jaderberg, M.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2016. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision* 116(1):1–20.

Jaderberg, M.; Vedaldi, A.; and Zisserman, A. 2014. Deep features for text spotting. In *European Conference on Computer Vision*, 512–528.

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 *IAPR International Conference on Document Analysis and Recognition*, 1484–1493.

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 *IAPR International Conference on Document Analysis and Recognition*, 1156–1160.

Kim, S.; Hori, T.; and Watanabe, S. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learn-

ing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 4835–4839.

Lee, C.-Y., and Osindero, S. 2016. Recursive recurrent nets with attention modeling for ocr in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2231–2239.

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*, 8610–8617.

Liao, M.; Zhang, J.; Wan, Z.; Xie, F.; Liang, J.; Lyu, P.; Yao, C.; and Bai, X. 2019. Scene text recognition from two-dimensional perspective. In *AAAI Conference on Artificial Intelligence*, 8714–8721.

Liu, Y.; Wang, Z.; Jin, H.; and Wassell, I. 2018. Synthetically supervised feature learning for scene text recognition. In *European Conference on Computer Vision*, 449–465.

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*, 7154–7161.

Long, S.; He, X.; and Yao, C. 2018. Scene text detection and recognition: The deep learning era. *CoRR* abs/1811.04256.

Long, J.; Shelhamer, E.; and Darrell, T. 2014. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 39(4):640–651.

Lucas, S. M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; and Young, R. 2003. ICDAR 2003 robust reading competitions. In *IAPR International Conference on Document Analysis and Recognition*, 682–687.

Luo, C.; Jin, L.; and Sun, Z. 2019. MORAN: A multi-object rectified attention network for scene text recognition. *Pattern Recognition* 90:109–118.

Luong, T.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 1412–1421.

Marti, U. V., and Bunke, H. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition* 5(1):39–46.

Mishra, A.; Alahari, K.; and Jawahar, C. 2012. Scene text recognition using higher order language priors. In *British Machine Vision Conference*, 1–11.

Neumann, L., and Matas, J. 2012. Real-time scene text localization and recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3538–3545.

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* 41(18):8027–8048.

Salvador, E. B.; Maria Jose, C. B.; Jorge, G. M.; and Francisco, Z. M. 2011. Improving offline handwritten text recognition with hybrid hmm/ann models. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(4):767–79.

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.* 39(11):2298–2304.

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*, 4168–4176.

Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; and Bai, X. 2018. ASTER: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.*

Sueiras, J.; Ruiz, V.; Sanchez, A.; and Velez, J. F. 2018. Offline continuous handwriting recognition using sequence to sequence neural networks. *Neurocomputing* 289:119–128.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Annual Conference on Neural Information Processing Systems*, 5998–6008.

Wang, K.; Babenko, B.; and Belongie, S. 2011. End-to-end scene text recognition. In *IEEE International Conference on Computer Vision*, 1457–1464.

Wang, T.; Wu, D. J.; Coates, A.; and Ng, A. Y. 2012. End-to-end text recognition with convolutional neural networks. In *International Conference on Pattern Recognition*, 3304–3308.

Xie, Z.; Huang, Y.; Zhu, Y.; Jin, L.; and Xie, L. 2019. Aggregation cross-entropy for sequence recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6538–6547.

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*, 3280–3286.

Zeiler, M. D. 2012. Adadelta: an adaptive learning rate method. *CoRR abs/1212.5701*.

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*, 2059–2068.

Zhang, Y.; Nie, S.; Liu, W.; Xu, X.; Zhang, D.; and Shen, H. T. 2019. Sequence-to-sequence domain adaptation network for robust text image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2740–2749.