Show, Attend and Read: A Simple and Strong Baseline for Irregular Text Recognition

Hui Li[1]*, Peng Wang[2]*, Chunhua Shen[1], Guyu Zhang[2]

[1]The University of Adelaide, and Australian Centre for Robotic Vision

[2]Northwestern Polytechnical University, China

* Indicates equal contribution

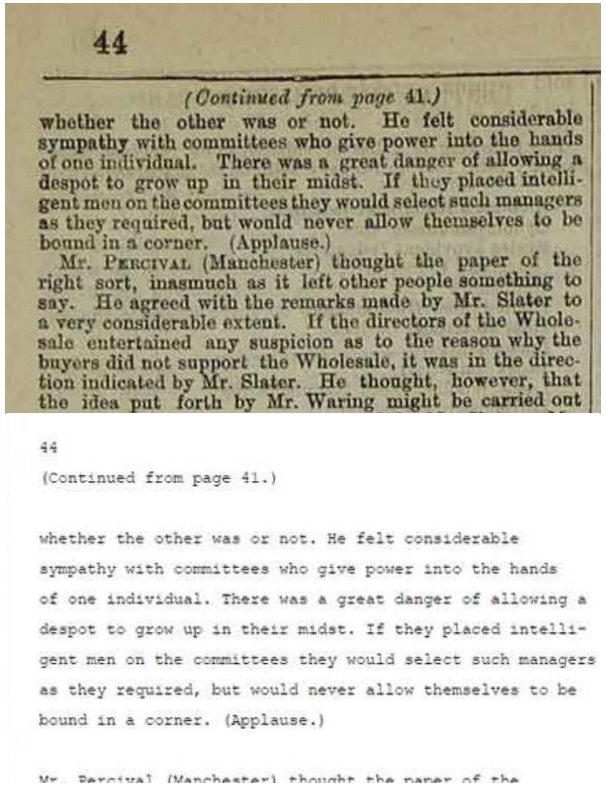
AAAI 2019 Pre-conference Presentation, Beijing

The Tasks of Text Recognition

OCR: Reading regular text in simple background

Reading regular text in natural scenes

Reading irregular text in natural scenes


44

(Continued from page 41.)

whether the other was or not. He felt considerable sympathy with committees who give power into the hands of one individual. There was a great danger of allowing a despot to grow up in their midst. If they placed intelligent men on the committees they would select such managers as they required, but would never allow themselves to be bound in a corner. (Applause.)

Mr. Percival (Manchester) thought the paper of the right sort, inasmuch as it left other people something to say. He agreed with the remarks made by Mr. Slater to a very considerable extent. If the directors of the Wholesale entertained any suspicion as to the reason why the buyers did not support the Wholesale, it was in the direction indicated by Mr. Slater. He thought, however, that the idea put forth by Mr. Waring might be carried out

44

(Continued from page 41.)

whether the other was or not. He felt considerable sympathy with committees who give power into the hands of one individual. There was a great danger of allowing a despot to grow up in their midst. If they placed intelligent men on the committees they would select such managers as they required, but would never allow themselves to be bound in a corner. (Applause.)

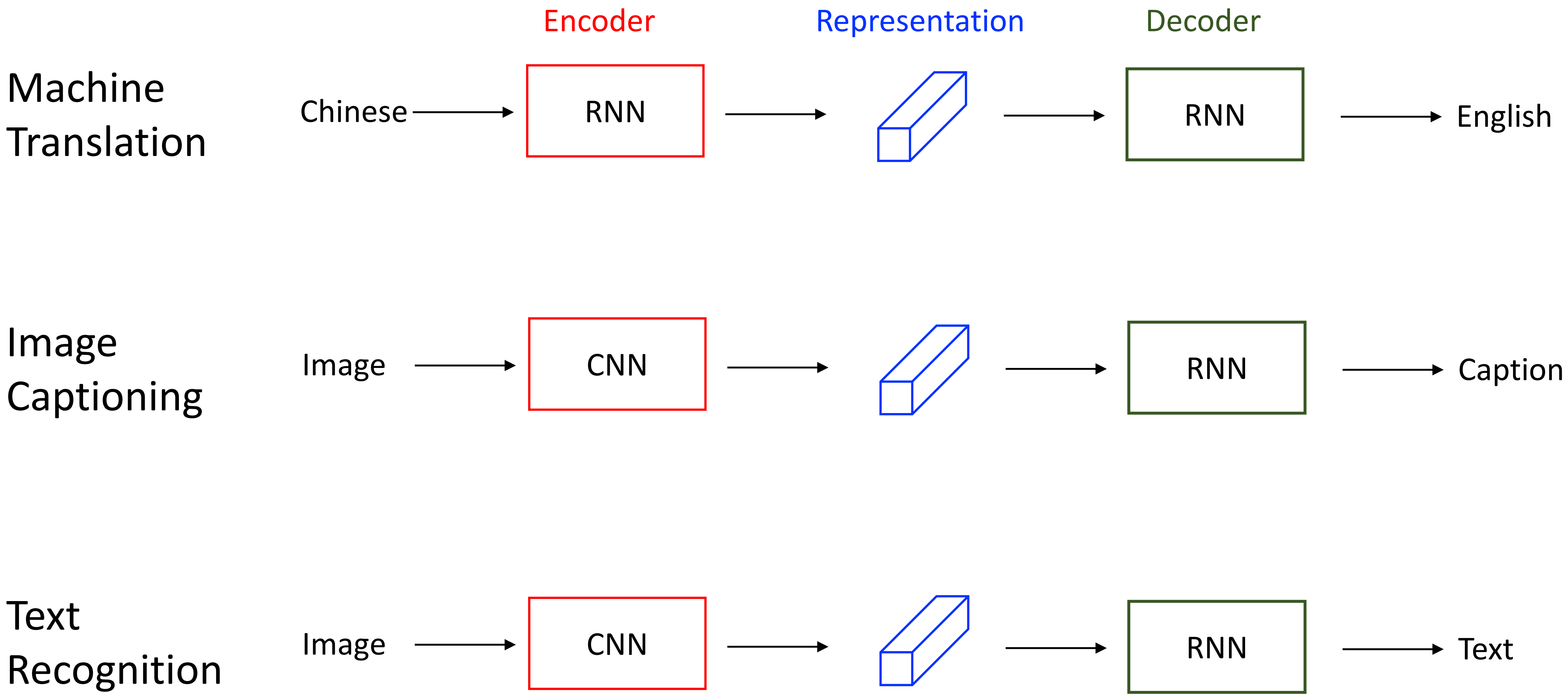
SCOTTISH

GRANDSTAN

HISTORIC

MANCHESTER

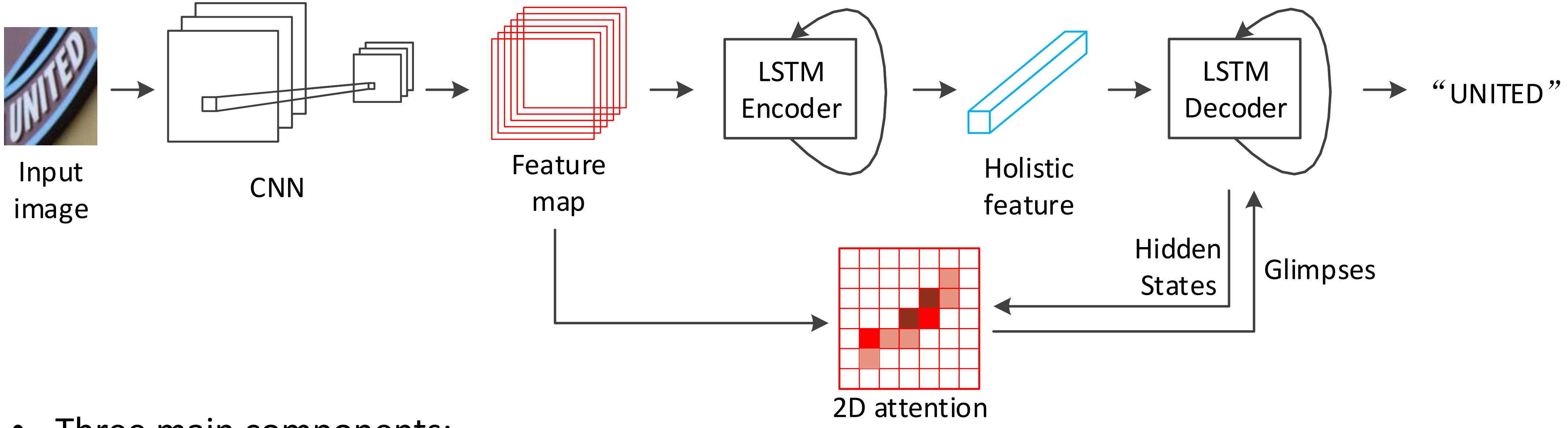
Relationship with Machine Translation



Existing Approaches

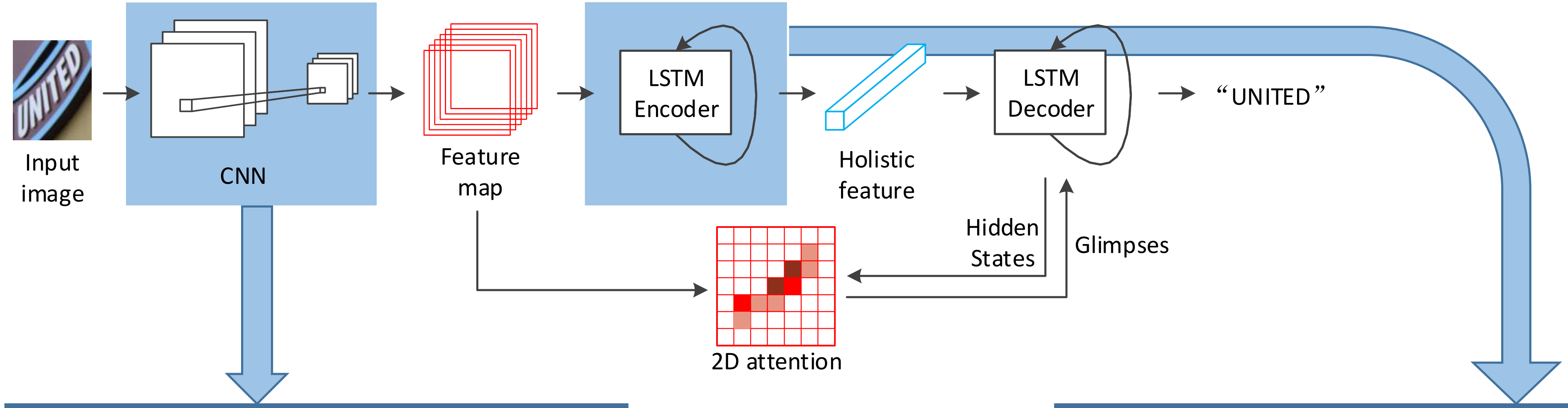
- Rectification based [Liu et al. 2016][Liu, Chen and Wong. 2018][Shi et al. 2018]
  - ✗ Difficult to tackle severe distortion or curvatures
- Attention based [Cheng et al. 2017]
  - ✗ Need character-level annotations which are hard to collect
- Multi-directional encoding based [Cheng et al. 2018]
  - ✗ sophisticated framework design and implementation

Overall framework

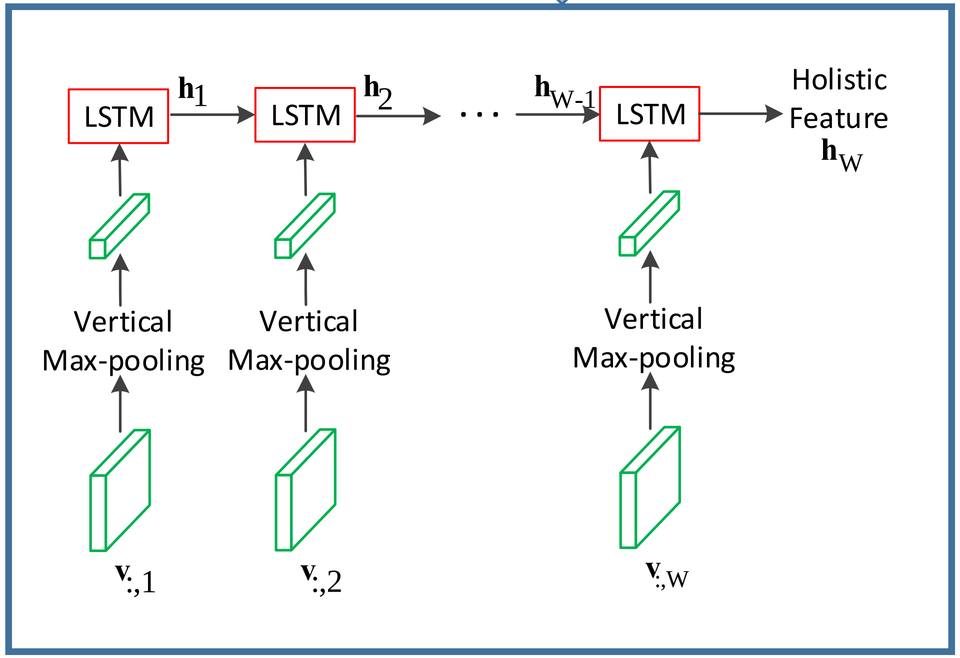


- Three main components:
  - ✓ A CNN+LSTM encoder
  - ✓ An LSTM decoder
  - ✓ A tailored 2D attention module
- Similar to the image captioning model "Show, Attend and Tell" [Xu et al., ICML, 2015]
- Easy to implement (main architecture in 100 lines)
- Require only word-level annotations
- State-of-the-art performance on both irregular and regular text recognition

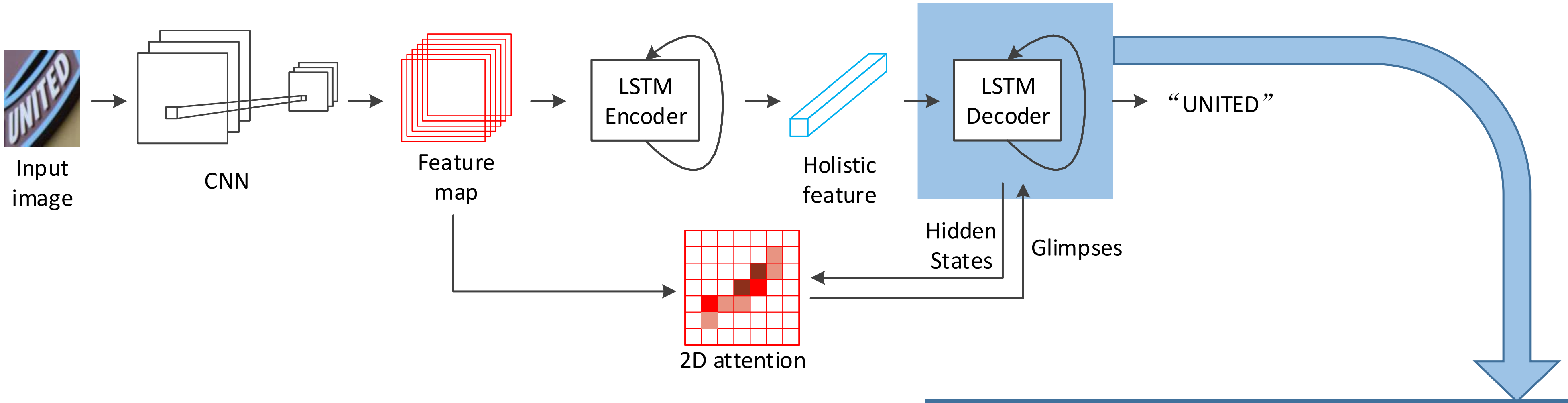
Overall framework: CNN+LSTM Encoder



Layer name	Configuration
Conv	$3 \times 3, 64$
Conv	$3 \times 3, 128$
Max-pooling	k:2×2, s:2×2
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 256 \\ Conv : 3 \times 3, 256 \end{bmatrix} \times 1$
Conv	$3 \times 3, 256$
Max-pooling	k:2×2, s:2×2
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 256 \\ Conv : 3 \times 3, 256 \end{bmatrix} \times 2$
Conv	$3 \times 3, 256$
Max-pooling	k:1×2, s:1×2
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 512 \\ Conv : 3 \times 3, 512 \end{bmatrix} \times 5$
Conv	$3 \times 3, 512$
Residual block	$\begin{bmatrix} Conv : 3 \times 3, 512 \\ Conv : 3 \times 3, 512 \end{bmatrix} \times 3$
Conv	$3 \times 3, 512$



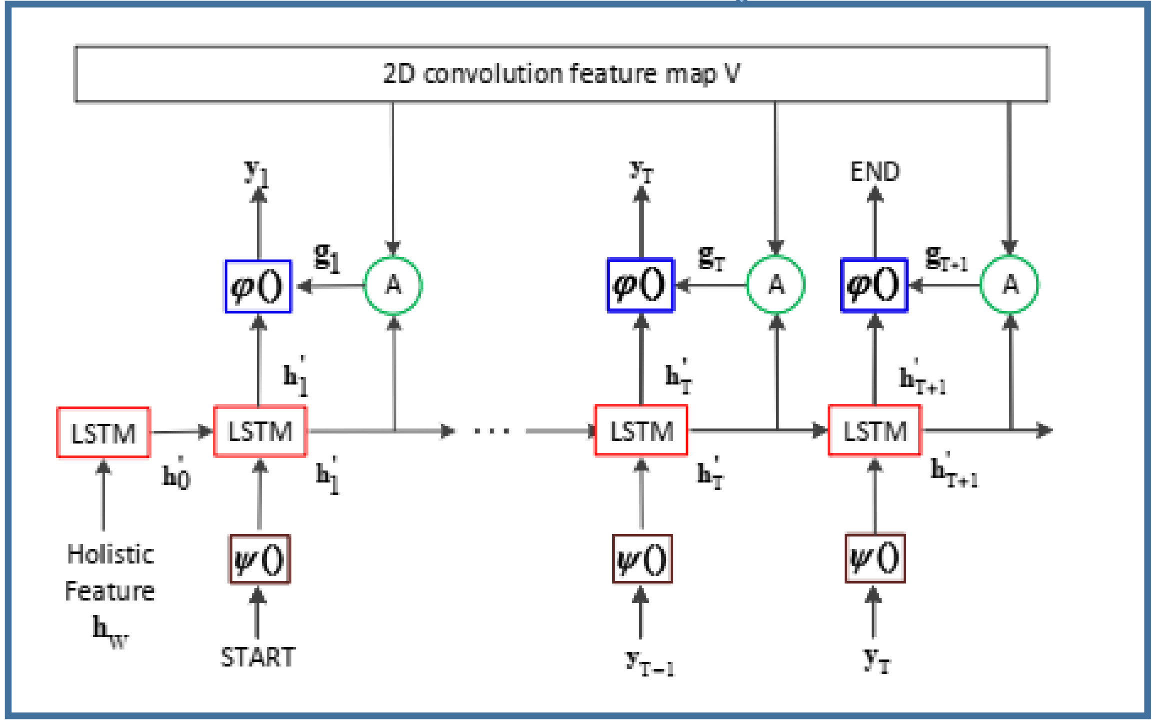
Overall framework: LSTM Decoder



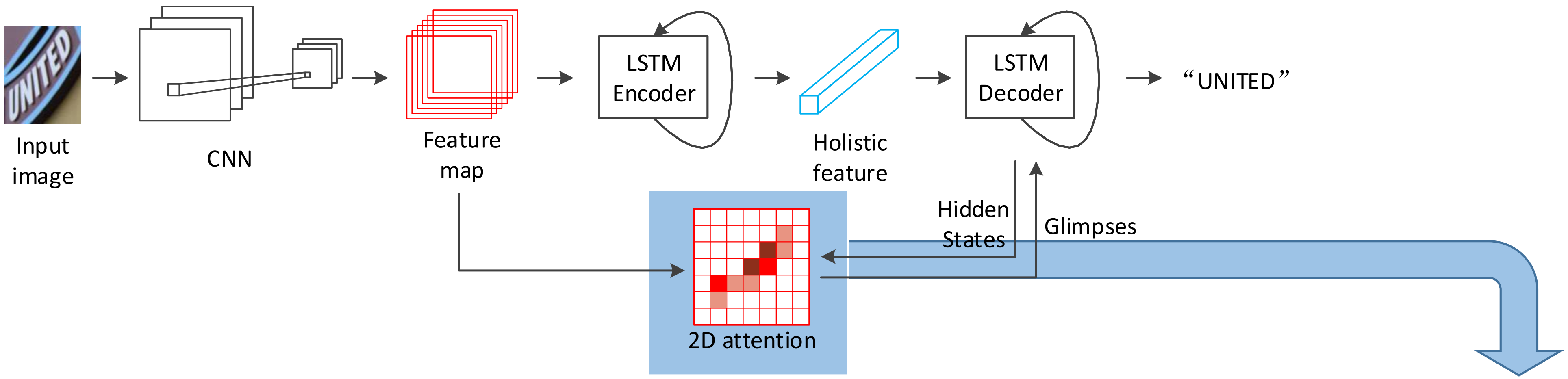
The decoder is an LSTM model with 2 layers and 512 hidden state size per layer. The output at timestep t is computed as:

$$\mathbf{y}_t = \varphi(\mathbf{h}'_t, \mathbf{g}_t) = \mathrm{softmax}(\mathbf{W}_o[\mathbf{h}'_t; \mathbf{g}_t])$$

where h'_t is the current hidden state and g_t is the output of the attention module.



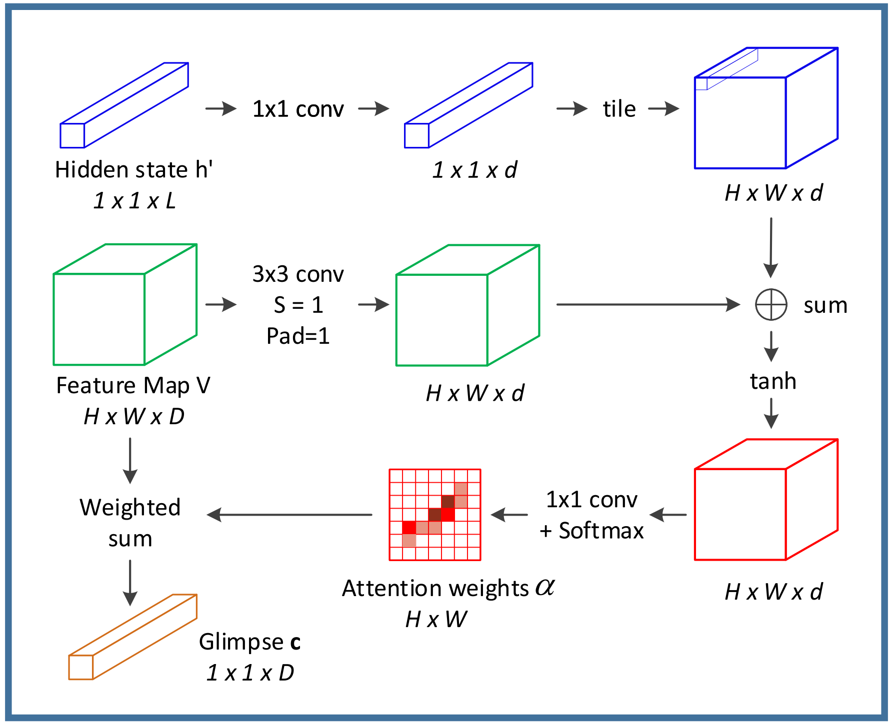
Overall framework: 2D Attention



In order to take neighborhood information into account, we propose a tailored 2D attention mechanism as follows:

$$\begin{cases} \mathbf{e}_{ij} = \tanh(\mathbf{W}_v \mathbf{v}_{ij} + \sum\limits_{p,q \in \mathcal{N}_{ij}} \tilde{\mathbf{W}}_{p-i,q-j} \cdot \mathbf{v}_{pq} + \mathbf{W}_h \mathbf{h}'_t), \\ \alpha_{ij} = \mathrm{softmax}(\mathbf{w}_e^T \cdot \mathbf{e}_{ij}), \\ \mathbf{g}_t = \sum\limits_{i,j} \alpha_{ij} \mathbf{v}_{ij}, \quad i = 1, \dots, H, \quad j = 1, \dots, W. \end{cases}$$

In which the computation can be accomplished by a series of convolution operations.



Experiments

- State-of-the-art word recognition performance especially for irregular scene text

Method	Regular Text						Irregular Text					
	IIIT5K			SVT		IC13	IC15	SVTP			CT80	COCO-T
	50	1k	None	50	None	None	None	50	Full	None	None	None
(Wang, Babenko, and Belongie 2011)	–	–	–	57.0	–	–	–	40.5	21.6	–	–	–
(Mishra, Alahari, and Jawahar 2012b)	64.1	57.5	–	73.2	–	–	–	45.7	24.7	–	–	–
(Phan et al. 2013)	–	–	–	73.7	–	–	–	75.6	67.0	–	–	
(Yao et al. 2014)	80.2	69.3	–	75.9	–	–	–	–	–	–	–	
(Jaderberg et al. 2015a)	97.1	92.7	–	95.4	80.7	90.8	–	–	–	–	42.7	
(He et al. 2016b)	94.0	91.5	–	93.5	–	–	–	–	–	–	–	
(Lee and Osindero 2016)	96.8	94.4	78.4	96.3	80.7	90.0	–	–	–	–	–	
(Wang and Hu 2017)	98.0	95.6	80.8	96.3	81.5	–	–	–	–	–	–	
(Shi et al. 2016)	96.2	93.8	81.9	95.5	81.9	88.6	–	91.2	77.4	71.8	59.2	–
(Liu et al. 2016)	97.7	94.5	83.3	95.5	83.6	89.1	–	*94.3*	83.6	73.5	–	
(Shi, Bai, and Yao 2017)	97.8	95.0	81.2	97.5	82.7	89.6	–	92.6	72.6	66.8	54.9	–
(Yang et al. 2017)*	97.8	96.1	–	95.2	–	–	–	93.0	80.2	75.8	69.3	–
(Cheng et al. 2017)*	99.3	97.5	87.4	97.1	85.9	93.3	70.6	92.6	81.6	71.5	63.9	–
(Liu et al. 2018)*	97.0	94.1	87.0	95.2	–	92.9	–	–	–	–	–	
(Liu, Chen, and Wong 2018)*	–	–	92.0	–	85.5	91.1	74.2	–	–	*78.9*	–	*59.3*
(Bai et al. 2018)*	*99.5*	97.9	88.3	96.6	87.5	**94.4**	73.9	–	–	–	–	
(Cheng et al. 2018)	**99.6**	98.1	87.0	96.0	82.8	–	68.2	94.0	*83.7*	73.0	76.8	
(Shi et al. 2018)	**99.6**	**98.8**	*93.4*	**99.2**	**93.6**	91.8	*76.1*	–	–	78.5	*79.5*	
SAR (Ours)	99.4	*98.2*	**95.0**	*98.5*	*91.2*	*94.0*	**78.8**	**95.8**	**91.2**	**86.4**	**89.6**	**66.8**

“50”, “1k” and “Full” represent lexicon sizes; “None” means lexicon-free; “*” indicates models trained with word-level and character-level annotations

Ablation Study

Model Configuration						IIIT5K	SVT	IC13	IC15	SVTP	CT80	COCO-T
Training data	CNN channels	Down-sampling ratio	Attention module	LSTM layers	Hidden state size							
Synth+Real	$\times 1$	$1/8, 1/4$	2D proposed	2	512	95.0	91.2	94.0	78.8	86.4	89.6	66.8
	$\times \mathbf{1/2}$	$1/8, 1/4$	2D proposed	2	512	92.7	88.7	92.0	75.6	81.3	86.8	62.6
	$\times 1$	$\mathbf{1/16}, 1/4$	2D proposed	2	512	93.8	90.3	92.7	77.4	84.5	89.2	64.8
	$\times 1$	$\mathbf{1/16}, \mathbf{1/8}$	2D proposed	2	512	94.0	90.6	93.1	76.2	83.7	87.5	63.7
	$\times 1$	$1/8, \mathbf{1/8}$	2D proposed	2	512	93.6	89.3	92.5	76.1	82.8	87.5	63.3
	$\times 1$	$1/8, 1/4$	**2D traditional**	2	512	94.0	90.1	92.3	77.2	84.3	87.5	64.2
	$\times 1$	$1/8, 1/4$	**1D**	2	512	93.0	89.9	90.2	76.6	83.6	84.7	65.4
	$\times 1$	$1/8, 1/4$	2D proposed	**1**	512	89.7	87.2	87.4	70.6	76.4	80.6	60.1
	$\times 1$	$1/8, 1/4$	2D proposed	2	**256**	94.0	89.3	92.8	76.8	83.7	86.5	63.8
OnlySynth	$\times 1$	$1/8, 1/4$	2D proposed	2	512	91.5	84.5	91.0	69.2	76.4	83.3	–

1. The volume of convolutional feature maps (#channels, width, height) should be sufficiently large to encode a rich variety of visual information for text recognition.

Ablation Study

Model Configuration						IIIT5K	SVT	IC13	IC15	SVTP	CT80	COCO-T
Training data	CNN channels	Down-sampling ratio	Attention module	LSTM layers	Hidden state size							
Synth+Real	×1	1/8, 1/4	2D proposed	2	512	95.0	91.2	94.0	78.8	86.4	89.6	66.8
	×1/2	1/8, 1/4	2D proposed	2	512	92.7	88.7	92.0	75.6	81.3	86.8	62.6
	×1	**1/16**, 1/4	2D proposed	2	512	93.8	90.3	92.7	77.4	84.5	89.2	64.8
	×1	**1/16, 1/8**	2D proposed	2	512	94.0	90.6	93.1	76.2	83.7	87.5	63.7
	×1	1/8, **1/8**	2D proposed	2	512	93.6	89.3	92.5	76.1	82.8	87.5	63.3
	×1	1/8, 1/4	**2D traditional**	2	512	94.0	90.1	92.3	77.2	84.3	87.5	64.2
	×1	1/8, 1/4	**1D**	2	512	93.0	89.9	90.2	76.6	83.6	84.7	65.4
	×1	1/8, 1/4	2D proposed	**1**	512	89.7	87.2	87.4	70.6	76.4	80.6	60.1
	×1	1/8, 1/4	2D proposed	2	**256**	94.0	89.3	92.8	76.8	83.7	86.5	63.8
OnlySynth	×1	1/8, 1/4	2D proposed	2	512	91.5	84.5	91.0	69.2	76.4	83.3	–

1. The volume of convolutional feature maps (#channels, width, height) should be sufficiently large to encode a rich variety of visual information for text recognition.

2. The tailored 2D attention module outperforms 1D attention and the traditional 2D attention modules.

Ablation Study

Model Configuration												
Training data	CNN channels	Down-sampling ratio	Attention module	LSTM layers	Hidden state size	IIIT5K	SVT	IC13	IC15	SVTP	CT80	COCO-T
Synth+Real	$\times 1$	$1/8$, $1/4$	2D proposed	2	512	95.0	91.2	94.0	78.8	86.4	89.6	66.8
	$\times \mathbf{1/2}$	$1/8$, $1/4$	2D proposed	2	512	92.7	88.7	92.0	75.6	81.3	86.8	62.6
	$\times 1$	$\mathbf{1/16}$, $1/4$	2D proposed	2	512	93.8	90.3	92.7	77.4	84.5	89.2	64.8
	$\times 1$	$\mathbf{1/16}$, $\mathbf{1/8}$	2D proposed	2	512	94.0	90.6	93.1	76.2	83.7	87.5	63.7
	$\times 1$	$1/8$, $\mathbf{1/8}$	2D proposed	2	512	93.6	89.3	92.5	76.1	82.8	87.5	63.3
	$\times 1$	$1/8$, $1/4$	**2D traditional**	2	512	94.0	90.1	92.3	77.2	84.3	87.5	64.2
	$\times 1$	$1/8$, $1/4$	**1D**	2	512	93.0	89.9	90.2	76.6	83.6	84.7	65.4
	$\times 1$	$1/8$, $1/4$	2D proposed	**1**	512	89.7	87.2	87.4	70.6	76.4	80.6	60.1
	$\times 1$	$1/8$, $1/4$	2D proposed	2	**256**	94.0	89.3	92.8	76.8	83.7	86.5	63.8
OnlySynth	$\times 1$	$1/8$, $1/4$	2D proposed	2	512	91.5	84.5	91.0	69.2	76.4	83.3	–

1. The volume of convolutional feature maps (#channels, width, height) should be sufficiently large to encode a rich variety of visual information for text recognition.
2. The tailored 2D attention module outperforms 1D attention and the traditional 2D attention modules.
3. The depth (#layers) and the width (hidden state size) should also be large enough

Ablation Study

Model Configuration												
Training data	CNN channels	Down-sampling ratio	Attention module	LSTM layers	Hidden state size	IIIT5K	SVT	IC13	IC15	SVTP	CT80	COCO-T
Synth+Real	$\times 1$	1/8, 1/4	2D proposed	2	512	95.0	91.2	94.0	78.8	86.4	89.6	66.8
	$\times \mathbf{1/2}$	1/8, 1/4	2D proposed	2	512	92.7	88.7	92.0	75.6	81.3	86.8	62.6
	$\times 1$	**1/16**, 1/4	2D proposed	2	512	93.8	90.3	92.7	77.4	84.5	89.2	64.8
	$\times 1$	**1/16**, **1/8**	2D proposed	2	512	94.0	90.6	93.1	76.2	83.7	87.5	63.7
	$\times 1$	1/8, **1/8**	2D proposed	2	512	93.6	89.3	92.5	76.1	82.8	87.5	63.3
	$\times 1$	1/8, 1/4	**2D traditional**	2	512	94.0	90.1	92.3	77.2	84.3	87.5	64.2
	$\times 1$	1/8, 1/4	**1D**	2	512	93.0	89.9	90.2	76.6	83.6	84.7	65.4
	$\times 1$	1/8, 1/4	2D proposed	**1**	512	89.7	87.2	87.4	70.6	76.4	80.6	60.1
	$\times 1$	1/8, 1/4	2D proposed	2	**256**	94.0	89.3	92.8	76.8	83.7	86.5	63.8
OnlySynth	$\times 1$	1/8, 1/4	2D proposed	2	512	91.5	84.5	91.0	69.2	76.4	83.3	–

1. The volume of convolutional feature maps (#channels, width, height) should be sufficiently large to encode a rich variety of visual information for text recognition.
2. The tailored 2D attention module outperforms 1D attention and the traditional 2D attention modules.
3. The depth (#layers) and the width (hidden state size) should also be large enough
4. Synthetic text images still cannot totally replace real images (COCO-Text). As real images usually do not come with character-level annotations, only models requiring word-level annotations can easily utilize these real training data.

Comparing with Rectification Based Methods

Rectification based methods:

- Design strategy: first rectify irregular text images to regular ones, and then recognize using 1d models
- Disadvantage: cannot tackle severe distortion or curvatures

Our approach:

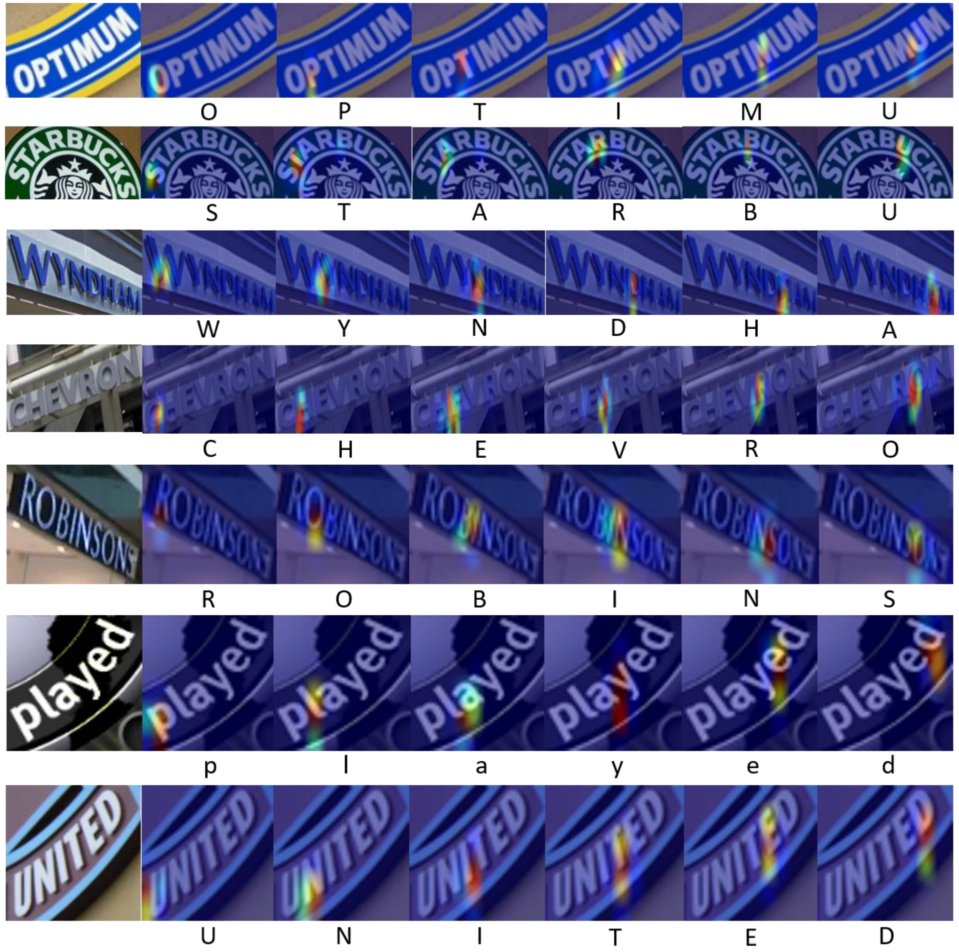
- Design strategy: directly recognize on the original images using 2d models
- Is capable of handling text of different layouts

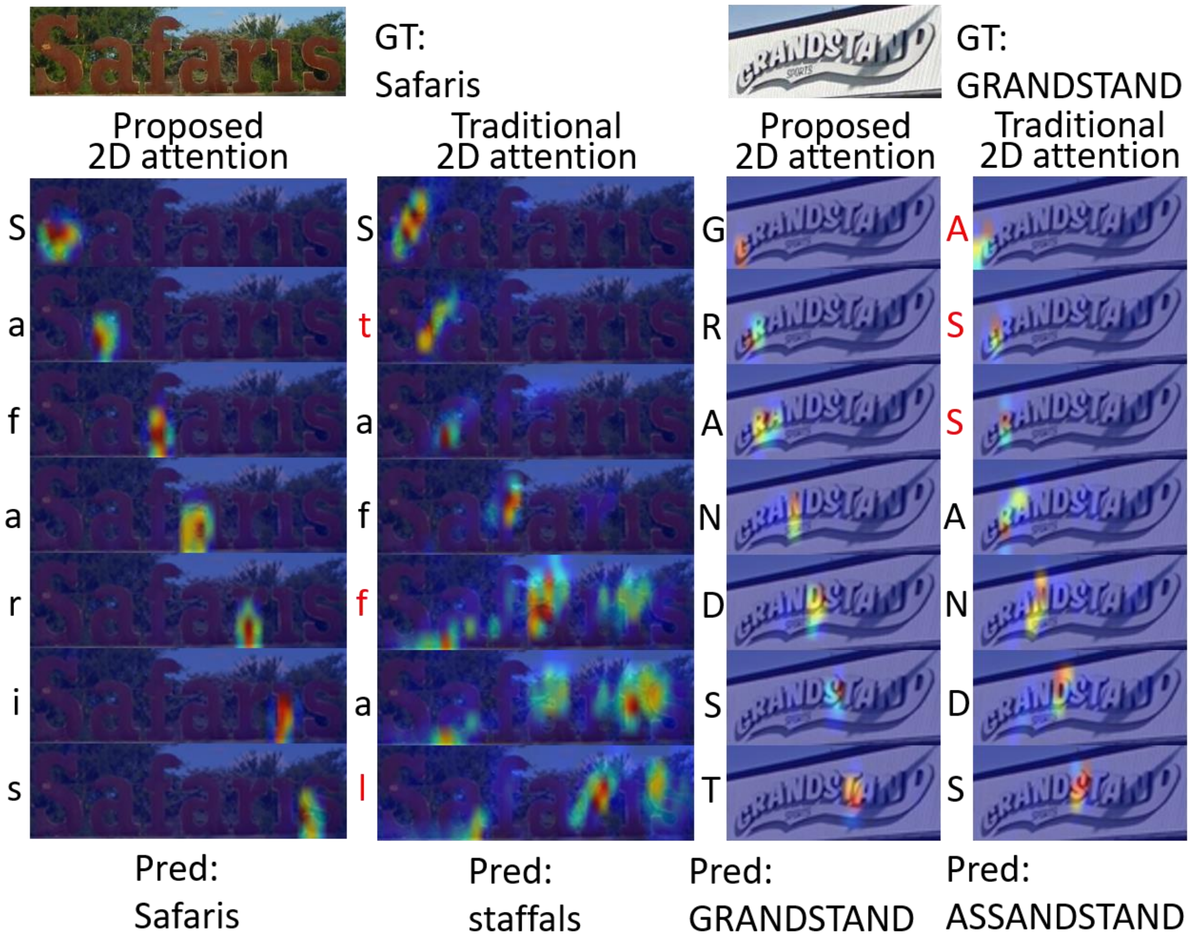
(a) Original Image (b) 2D attention (ours) (c) Rectification (Shi et al. 2018)

Visualization of 2D Attention Weights

- Our 2D attention model can be trained to approximately localize characters without character-level annotations (a semi-supervised fashion).

- The proposed 2D attention model shows more accurate localization and better recognition results, compared with the traditional 2D attention model.





Failure Cases

There are a variety of reasons for failure, such as:

- blurry
- partial occlusion
- extreme distortion
- uneven lighting condition
- uncommon fonts
- vertical text

GT: RAFFLES
Pred: CAFE

GT: TAGHeuer
Pred: TALKEUER

GT: TOWN
Pred: Titutt

GT: SWAROVSKI
Pred: SEAROVILI

GT: Glaillo
Pred: GLOSFILLE

GT: WAREHOUSE
Pred: MUCOIMUDS

GT: H
Pred: HALL

GT: D
Pred: SOT

GT: HOT
Pred: HONDA

GT: Expert
Pred: Capert

Thanks

Source code:

https://github.com/wangpengnorman/SAR-Strong-Baseline-for-Text-Recognition