MEAN: Multi-Element Attention Network for Scene Text Recognition

Ruijie Yan[1], Liangrui Peng [1], Shanyu Xiao[1], Gang Yao [1], and Jaesik Min [2]

Presented by Ruijie Yan

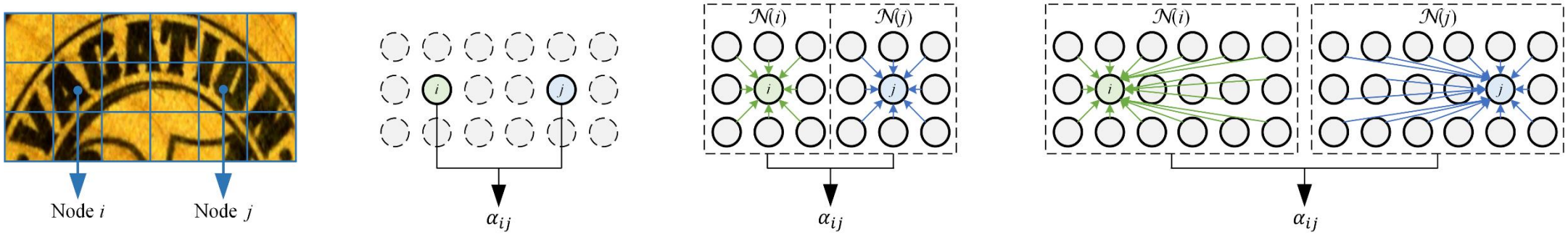
yrj17@mails.tsinghua.edu.cn

[1] Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, China

[2] Hyundai Motor Group AIRS Company, Seoul, Korea

Motivation

- Challenges of scene text recognition
  - How to handle wide variances in styles, orientations, and image qualities
  - How to sufficiently explore 2D spatial information
- Our idea: Multi-Element Attention (MEA)
  - Incorporating graph structure modeling into self-attention mechanism [1]
  - Assigning various adjacency matrices to the graph



(a) Graph structure modeling (b) MEA-Local (c) MEA-Neighbor (d) MEA-Global

Fig. 1. (a) Features extracted from an input image is modeled as an undirected graph; (b)-(d) Schematic diagram of the attention weights computation of three different MEAs.

Multi-Element Attention

- Self-attention mechanism

$$SA(X) = \phi\left(\frac{1}{\sqrt{d}}(XW_Q)\cdot(XW_K)\right)XW_V$$

- MEA is a generalized form of the self-attention mechanism

$$MEA(X) = \phi\left(\frac{1}{\sqrt{d}}(AXW_Q)\cdot(BXW_K)\right)XW_V$$

- Three different implementations of AXW_Q and BXW_K
  - MEA-Local: 1×1 convolutions with **local** receptive field
  - MEA-Neighbor: $m \times n$ convolutions with **neighbor** receptive field
  - MEA-Global: graph convolutions with **global** receptive field

Multi-Element Attention Network (MEAN)

- MEAN consists of a CNN, an encoder, and a decoder
  - CNN: a modified EfficientNet-B3[2] with U-shaped structure
  - Encoder: three types of MEA mechanism
  - Decoder: Transformer decoder

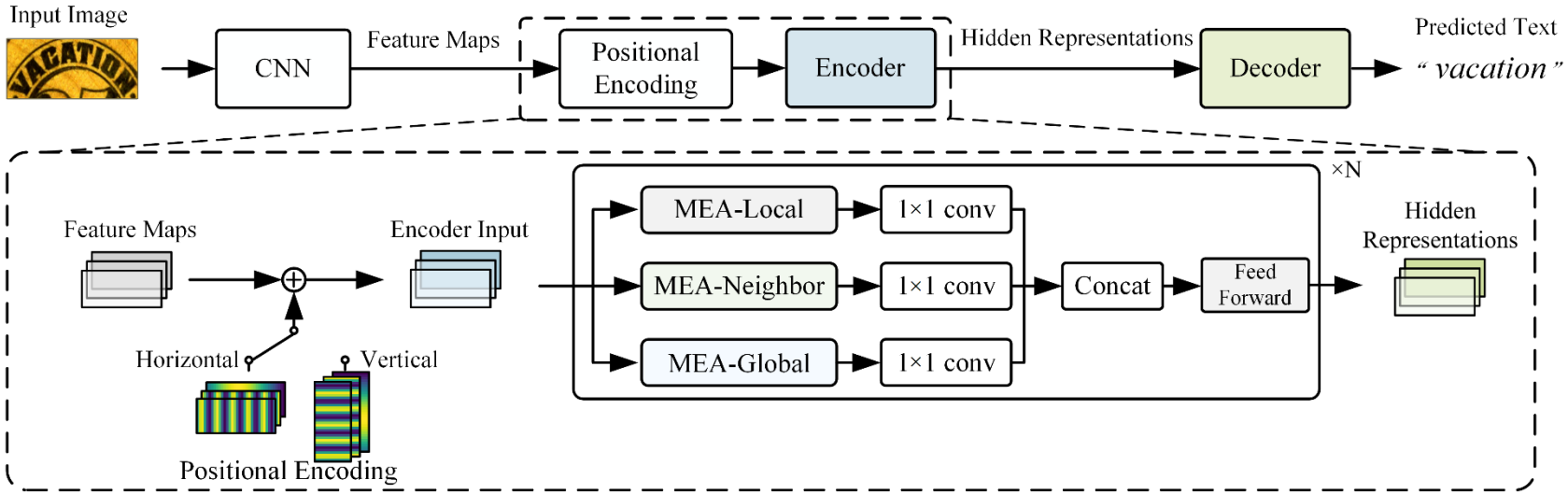


Fig. 2. System framework of MEAN that consists of: a CNN, an encoder equipped with the MEA mechanism, and a decoder. Orientational positional encoding is added into features maps output by the CNN.

Multi-Element Attention Network (MEAN)

- Orientational positional encoding: handling multi-oriented text images

Horizontal text images

$$PE^{H}_{(i,j,2k)} = \sin(j/L^{2k/d})$$
$$PE^{H}_{(i,j,2k+1)} = \cos(j/L^{2k/d})$$

Vertical text images

$$PE^{V}_{(i,j,2k)} = \sin(i/L^{2k/d})$$
$$PE^{V}_{(i,j,2k+1)} = \cos(i/L^{2k/d})$$

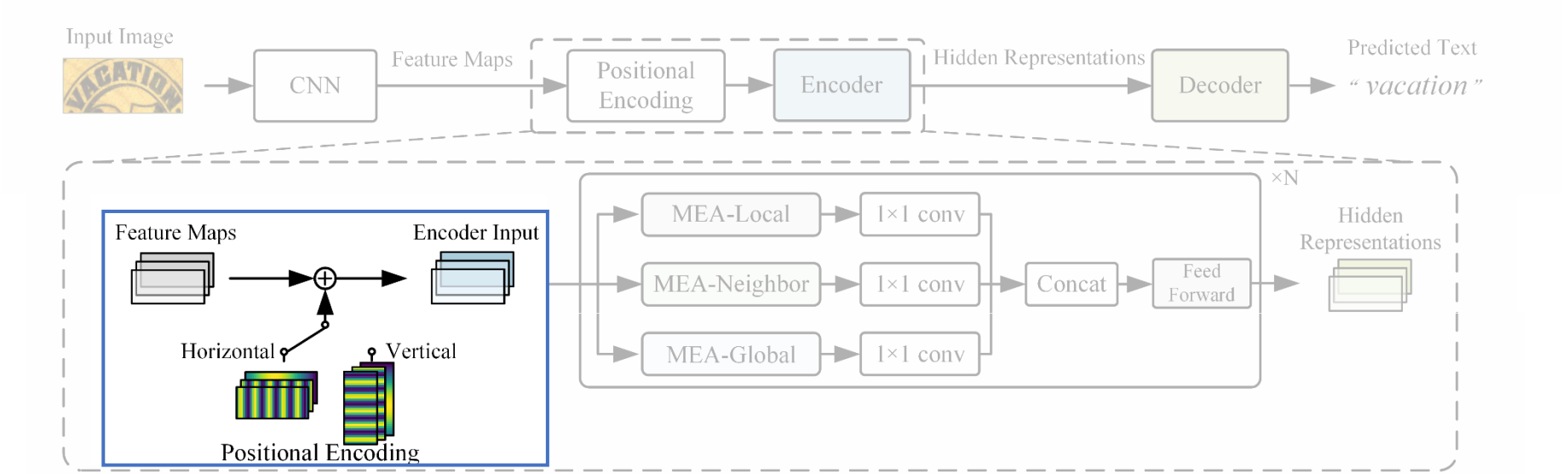


Fig. 2. System framework of MEAN that consists of: a CNN, an encoder equipped with the MEA mechanism, and a decoder. Orientational positional encoding is added into features maps output by the CNN.

Experiments

- English scene text recognition
  - Comparing with previous state-of-the-art methods
  - Training set: MJSynth, SynthText
  - Test set: IIIT5k, SVT, IC03, IC13, IC15, SVTP, CUTE

- Chinese scene text recognition
  - Exploring the performance of recognizing multi-oriented texts
  - Training set: self-synthesized samples, a subset of RCTW
  - Test set: a subset of RCTW

English Scene Text Recognition

Table 1. Word recognition accuracy (%) across methods and datasets. MJ, ST, Char, and Add denote MJSynth, SynthText, character bounding boxes, and additional training data, respectively. The best results are marked in **bold**.

Model	Training data	Regular text datasets				Irregular text datasets		
		IIIT5k	SVT	IC03	IC13	IC15	SVTP	CUTE
FAN (Cheng et al.) [3]	MJ+ST+Char	87.4	85.9	94.2	93.3	70.6	-	-
Mask TextSpotter (Liao et al.) [4]	MJ+ST+Char	95.3	91.8	95.0	95.3	78.2	83.6	88.5
SAR (Li et al.) [5]	MJ+ST+Add	95.0	91.2	-	94.0	78.8	86.4	**89.6**
AON (Cheng et al.) [6]	MJ+ST	87.0	82.8	91.5	-	68.2	73.0	76.8
EP (Bai et al.) [7]	MJ+ST	88.3	87.5	94.6	94.4	73.9	-	-
ACE (Xie et al.) [8]	MJ+ST	82.3	82.6	92.1	89.7	68.9	70.1	82.6
MORAN (Luo et al.) [9]	MJ+ST	91.2	88.3	95.0	92.4	68.8	76.1	77.4
DAN (Wang et al.) [10]	MJ+ST	94.3	89.2	95.0	93.9	74.5	80.0	84.4
ASTER (Shi et al.) [11]	MJ+ST	93.4	89.5	94.5	91.8	76.1	78.5	79.5
SRN (Yu et al.) [12]	MJ+ST	94.8	91.5	-	**95.5**	**82.7**	85.1	87.8
MEAN	MJ+ST	**95.9**	**94.3**	**95.9**	95.1	79.7	**86.8**	87.2

Multi-Oriented Chinese Scene Text Recognition

- Baseline is a CNN-Transformer network with 1D attention mechanism
- Trained on only horizontal or vertical text images
  - MEAN achieves a slightly higher accuracy than baseline
- Trained on both horizontal and vertical text images
  - Performance of baseline is significantly degraded
  - MEAN achieves even higher performance

Table 2. Word accuracy (%) of different models for multi-oriented Chinese scene text recognition. "H", "V", and "H & V" denote the model is trained on only horizontal text images, only vertical text images, and both horizontal and vertical text images.

Test set	Baseline			MEAN		
	H	V	H & V	H	V	H & V
Horizontal	74.2	-	52.2	**77.4**	-	**81.6**
Vertical	-	74.6	36.0	-	**78.6**	**86.0**

Effectiveness of MEA

- MEAN-Neighbor and MEAN-Global outperform MEAN-Local
- MEAN with all three types of MEAs achieves the highest performance

Table 3. Word accuracy (%) of models with different variants of MEAs.

Model	#params		English		Chinese	
	English	Chinese	SVT	SVTP	Horizontal	Vertical
MEAN-Local	23.4M	29.9M	93.5	86.5	80.2	83.2
MEAN-Neighbor	28.1M	34.6M	93.8	**86.8**	81.2	83.8
MEAN-Global	23.7M	30.1M	93.5	85.9	80.8	84.8
MEAN	31.0M	37.5M	**94.3**	**86.8**	**81.6**	**86.0**

Visualization of attention weights

- MEA-Neighbor and MEAN-Global focus more on foreground areas
- Three types of MEAs are complementary

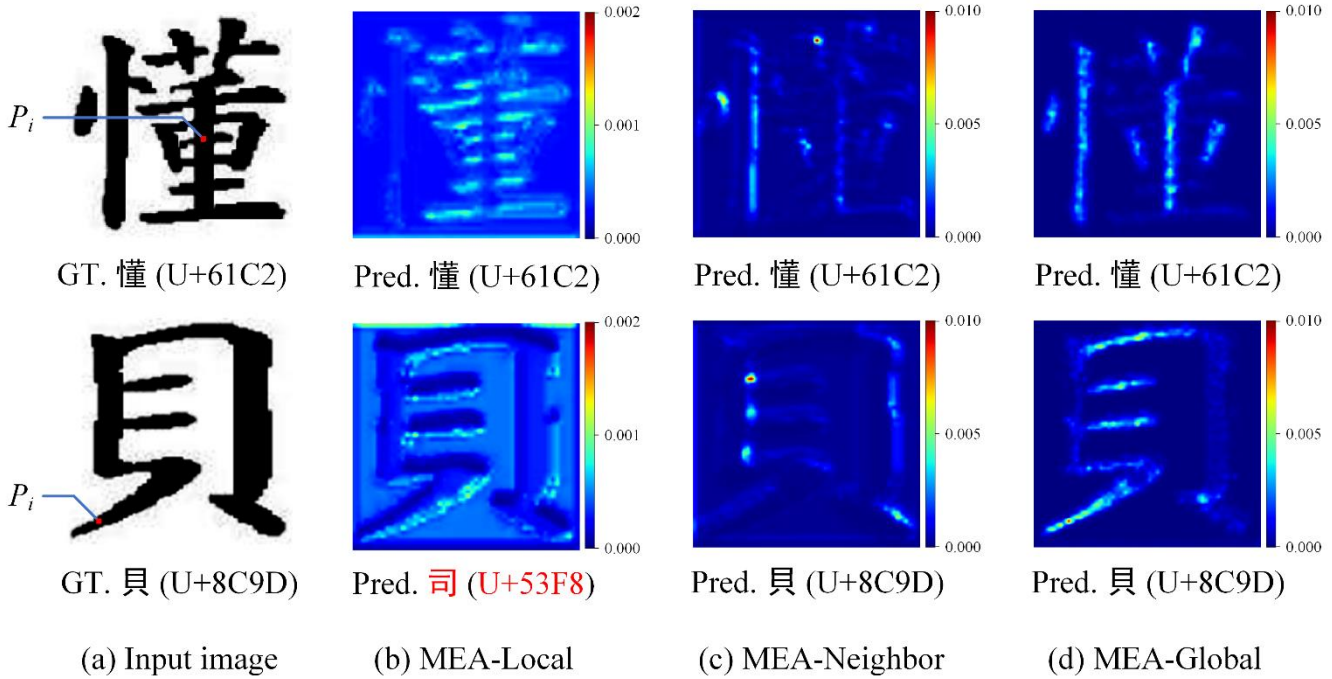


Fig. 3. Visualization of attention weights α_{ij}. (a) Input image, and the red point denotes position i. (b)-(d) The attention weights of MEA-Local, MEA-Neighbor, and MEA-Global, respectively.

Recognition Examples

GT: ronaldo
Output: ronaldo

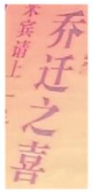
GT: allahabad
Output: allahabad

GT: starbucks
Output: starbucks

GT: 富观街
(u+ 5bcc 89c2 8857)
Output: 富观街
(u+ 5bcc 89c2 8857)

GT: 从我做起
(u+ 4ece 6211 505a 8d77)
Output: 从我做起
(u+ 4ece 6211 505a 8d77)

GT: 保护环境
(u+ 4fdd 62a4 73af 5883)
Output: 保护环境
(u+ 4fdd 62a4 73af 5883)



GT: 乔迁之喜
(u+ 4e54 8fc1 4e4b 559c)
Output: 乔迁之喜
(u+ 4e54 8fc1 4e4b 559c)

GT: 鸟语花香
(u+ 9e1f 8bed 82b1 9999)
Output: 鸟语花香
(u+ 9e1f 8bed 82b1 9999)

GT: 沙县小吃
(u+ 6c99 53bf 5c0f 5403)
Output: 沙县小吃
(u+ 6c99 53bf 5c0f 5403)

- Support for curved, skewed, and multi-oriented texts

Reference

[1] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in NeurIPS, 2017, pp. 5998–6008.

[2] M. Tan, Q. V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks,” in ICML, 2019, pp. 6105–6114.

[3] Z. Cheng, F. Bai, Y. Xu et al., “Focusing attention: Towards accurate text recognition in natural images,” in ICCV, 2017, pp. 5076–5084.

[4] M. Liao, P. Lyu, M. He et al., “Mask Textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes,” in ECCV, 2018, pp. 67–83.

[5] H. Li, P. Wang, C. Shen, and G. Zhang, “Show, attend and read: A simple and strong baseline for irregular text recognition,” in AAAI, 2019, pp. 8610–8617.

[6] Z. Cheng, Y. Xu, F. Bai et al., “AON: Towards arbitrarily-oriented text recognition,” in CVPR, 2018, pp. 5571–5579.

[7] F. Bai, Z. Cheng, Y. Niu et al., “Edit probability for scene text recognition,” in CVPR, 2018, pp. 1508–1516.

[8] Z. Xie, Y. Huang, Y. Zhu et al., “Aggregation cross-entropy for sequence recognition,” in CVPR, 2019, pp. 6538–6547.

[9] C. Luo, L. Jin, and Z. Sun, “MORAN: A multi-object rectified attention network for scene text recognition,” Pattern Recognition, vol. 90, pp. 109–118, 2019.

[10] T. Wang, Y. Zhu, L. Jin et al., “Decoupled attention network for text recognition,” in AAAI, 2020, pp. 12216–12224.

[11] B. Shi, M. Yang, X. Wang et al., “ASTER: An attentional scene text recognizer with flexible rectification,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2035–2048, 2019.

[12] D. Yu, X. Li, C. Zhang et al., “Towards accurate scene text recognition with semantic reasoning networks,” in CVPR, 2020, pp. 12113–12122.


thank you