MEAN: Multi-Element Attention Network for Scene Text Recognition

Ruijie Yan*, Liangrui Peng*, Shanyu Xiao*, Gang Yao*, and Jaesik Min†

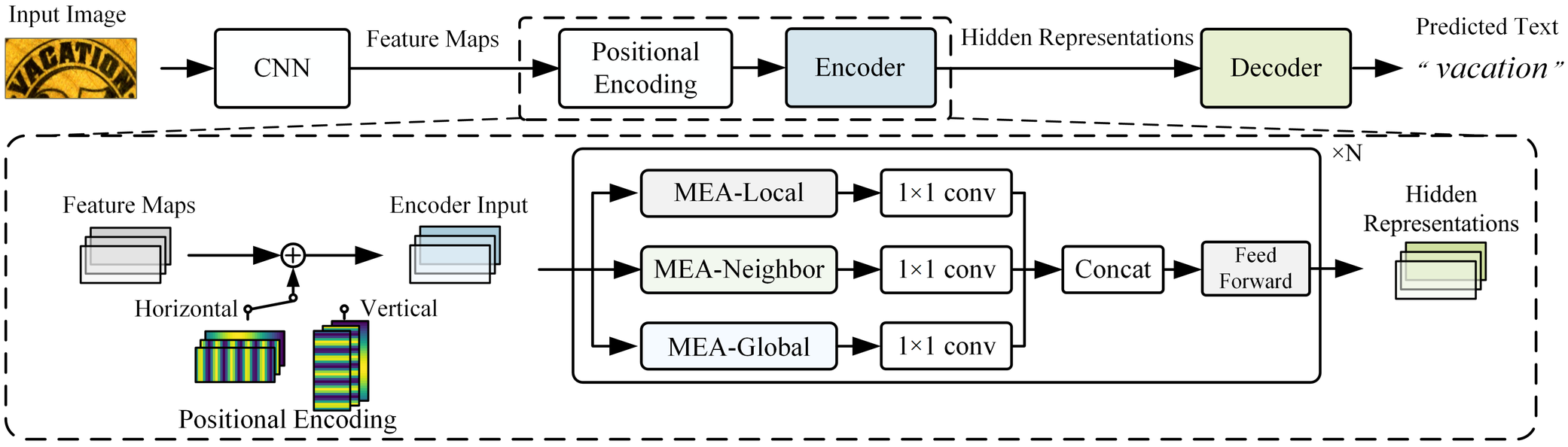
*Beijing National Research Center for Information Science and Technology

Dept. of Electronic Engineering, Tsinghua University, Beijing, China

† Hyundai Motor Group AIRS Company, Seoul, Korea

Introduction

We propose a novel multi-element attention (MEA) mechanism to exploit geometric structures from local to global levels in feature maps extracted from a scene text image. A multi-element attention network (MEAN) is constructed by using three types of MEAs. Experiments on English and Chinese scene text recognition experiments demonstrates the effectiveness of MEAN on regular, irregular, and multi-oriented texts.

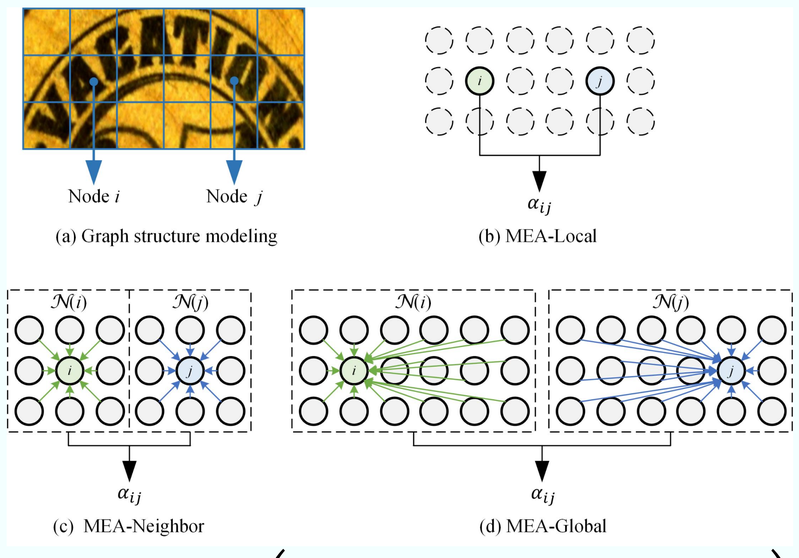


System framework of MEAN that consists of a CNN, an encoder, and a decoder.

Method

Multi-Element Attention (MEA)

- **Idea**
  - Incorporating graph structure modeling into self-attention mechanism and assigning various adjacency matrices to the graph.



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

- **Three implementations of AXW_Q and BXW_K**
  - **MEA-Local**: 1×1 convolutions
  - **MEA-Neighbor**: $m\times n$ convolutions
  - **MEA-Global**: graph convolutions

Multi-Element Attention Network (MEAN)

- **Framework**
  - MEAN consists of a U-shaped CNN, an encoder with three MEAs, and a decoder.
- **Orientational positional encoding**

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

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

Experiment

- **English scene text recognition**

Word Accuracy (%) across different models and datasets

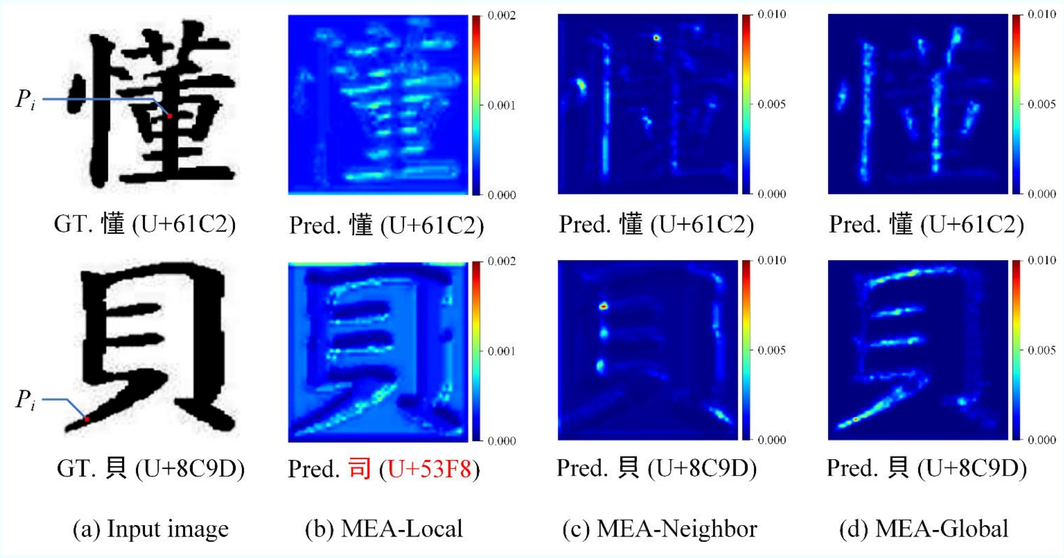
Model	IIIT5k	SVT	IC03	IC13	IC15	SVTP	CUTE
Mask TextSpotter	95.3	91.8	95.0	95.3	78.2	83.6	88.5
SAR	95.0	91.2	-	94.0	78.8	86.4	**89.6**
ASTER	93.4	89.5	94.5	91.8	76.1	78.5	79.5
SRN	94.8	91.5	-	**95.5**	**82.7**	85.1	87.8
MEAN	**95.9**	**94.3**	**95.9**	95.1	79.7	**86.8**	87.2

- **Multi-oriented Chinese scene text recognition**

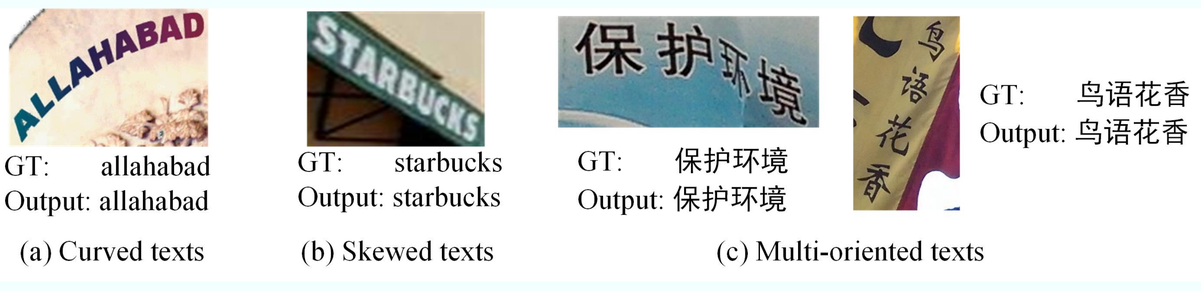
Word Accuracy (%) of different models on multi-oriented texts

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**

- **Visualization of attention scores**



- **Recognition Examples**



Acknowledgment:

- **This research is supported by a joint research project between Hyundai Motor Group AIRS Company and Tsinghua University.**
- **The second author is partially supported by National Key R&D Program of China and a grant from the Institute for Guo Qiang, Tsinghua University.**