MEAN: Multi-Element Attention Network for Scene Text Recognition

Ruijie Yan\textsuperscript{1}, Liangrui Peng\textsuperscript{1}, Shanyu Xiao\textsuperscript{1}, Gang Yao\textsuperscript{1}, and Jaesik Min\textsuperscript{2}

Presented by Ruijie Yan
yrj17@mails.tsinghua.edu.cn

\textsuperscript{1} Beijing National Research Center for Information Science and Technology
Department of Electronic Engineering, Tsinghua University, Beijing, China

\textsuperscript{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

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: \( 1 \times 1 \) 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
  - 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**.

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

<table>
<thead>
<tr>
<th>Test set</th>
<th>Baseline</th>
<th>MEAN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>H</td>
<td>V</td>
</tr>
<tr>
<td>Horizontal</td>
<td>74.2</td>
<td>-</td>
</tr>
<tr>
<td>Vertical</td>
<td>-</td>
<td>74.6</td>
</tr>
</tbody>
</table>
Effectiveness of MEA

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

<table>
<thead>
<tr>
<th>Model</th>
<th>#params</th>
<th>English</th>
<th>Chinese</th>
<th>Chinese</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>English</td>
<td>SVT</td>
<td>SVTP</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chinese</td>
<td>Horizontal</td>
<td>Vertical</td>
</tr>
<tr>
<td>MEAN-Local</td>
<td>23.4M</td>
<td>29.9M</td>
<td>93.5</td>
<td>86.5</td>
</tr>
<tr>
<td>MEAN-Neighbor</td>
<td>28.1M</td>
<td>34.6M</td>
<td>93.8</td>
<td><strong>86.8</strong></td>
</tr>
<tr>
<td>MEAN-Global</td>
<td>23.7M</td>
<td>30.1M</td>
<td>93.5</td>
<td>85.9</td>
</tr>
<tr>
<td>MEAN</td>
<td>31.0M</td>
<td>37.5M</td>
<td><strong>94.3</strong></td>
<td><strong>86.8</strong></td>
</tr>
</tbody>
</table>

Table 3. Word accuracy (%) of models with different variants of MEAs.
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 $\alpha_{ij}$. (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

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

thank you