SHAPE: Shifted Absolute Position Embedding for Transformers

Shun Kiyono, Sosuke Kobayashi, Jun Suzuki, Kentaro Inui

Overview: building absolute position embedding that can extrapolate and shift-invariant

Transformer

(b) Relative Position Embedding (RPE)

(a) Absolute Position Embedding (APE)

(c) Shifted APE (SHAPE)

i.e. generalize to lengths unseen during training

Summary of Position Representations

<table>
<thead>
<tr>
<th>Performance</th>
<th>APE</th>
<th>RPE</th>
<th>SHAPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed</td>
<td>😞</td>
<td>😞</td>
<td>😞</td>
</tr>
<tr>
<td>Implementation</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

SHAPE Implementation: a single line

pos_emb = self.training × torch.randn(0, pos_idx) + 1

Random shift in APE can prevent the use of absolute positions
→ Model learns shift invariance?

Idea: Bring RPE’s success to APE

Shift invariance: spatial shift does not change function’s output

\( (Key, Value)_{t-1} \) is same as \( (Key, Value)_{t} \)

Experiment on Machine Translation (WMT EnDe)

① Vanilla

- Standard setting for MT
- Sanity check of baseline performance

② Extrapolate

- Remove sequence longer than 50 tokens
- Evaluate performance on unseen length

③ Interpolate

- Concatenate neighboring sequences
- Tokens are more infrequent at given position

Experimental Result and Observations

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Model</th>
<th>Valid</th>
<th>Test</th>
<th>Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>APE†</td>
<td>23.61</td>
<td>30.46</td>
<td>x1.00</td>
</tr>
<tr>
<td></td>
<td>RPE†</td>
<td>23.67</td>
<td>30.54</td>
<td>x0.91</td>
</tr>
<tr>
<td></td>
<td>SHAPE †</td>
<td>23.63</td>
<td>30.49</td>
<td>x1.01</td>
</tr>
<tr>
<td>Extrapolate</td>
<td>APE</td>
<td>22.18</td>
<td>29.22</td>
<td>x1.00</td>
</tr>
<tr>
<td></td>
<td>RPE</td>
<td>22.97</td>
<td>29.86</td>
<td>x0.91</td>
</tr>
<tr>
<td></td>
<td>SHAPE</td>
<td>22.96</td>
<td>29.80</td>
<td>x0.99</td>
</tr>
<tr>
<td>Interpolate</td>
<td>APE</td>
<td>31.40</td>
<td>38.23</td>
<td>x1.00</td>
</tr>
<tr>
<td></td>
<td>RPE†</td>
<td>30.46</td>
<td>37.93</td>
<td>x1.00</td>
</tr>
<tr>
<td></td>
<td>SHAPE</td>
<td>32.50</td>
<td>39.09</td>
<td>x0.99</td>
</tr>
</tbody>
</table>

- All models are comparable
- SHAPE has no risk of performance drop
- Both RPE and SHAPE outperform APE
- SHAPE is as fast as APE while RPE is not
- RPE is prohibitively slow
- SHAPE outperforms APE

Analysis 1: SHAPE is Shift-invariant

Cosine Similarities of Encoder Hidden States

Analysis 2: SHAPE Extrapolates better than APE

- SHAPE outperforms APE in gray no-training-data region
- Better extrapolation than APE