Rethinking and Improving Relative Position Encoding for Vision Transformer

Team E: Avirup Ghosh, Shiyi Gong, Shenyang Liu, Rishabh Patel

Published 2021 IEEE/CVF
Contents

• Background and previous work
• Four new relative position encoding methods
• Implementation details and Analysis on Relative Position Encoding
• Comparison on image classification and object detection and conclusion
Background and previous work

- Self-Attention
- Absolute Position Encoding and Relative Position Encoding (RPE)
- Shaw’s RPE
- RPE in Transformer-XL
- Huang’s RPE
- RPE in SASA
- RPE in Axial-Deeplab
Self-attention

\[ z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^V). \quad (1) \]

\[ \alpha_{ij} = \frac{\text{softmax}(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})}. \quad (2) \]

\[ e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T}{\sqrt{d_z}}. \quad (3) \]
Absolute Position Encoding and Relative Position Encoding (RPE)

- Absolute Positional encoding $\mathbf{p} = (p_1, \ldots, p_n)$
  
  $x_i = x_i + p_i$, 

- Relative Position Encoding (RPE)

  \[ z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^V + p_{ij}^V), \]  

  \[ \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})}, \]  

  \[ e_{ij} = \frac{(x_i W^Q + p_{ij}^Q) (x_j W^K + p_{ij}^K)^T}{\sqrt{d_z}}. \]
Shaw’s RPE

\[ \text{clip}(x, k) = \max(-k, \min(k, x)), \]

\[ z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^V + p_{V(i-j,k)}^V), \]

\[ e_{ij} = \frac{(x_i W^Q)(x_j W^K + p_{K(i-j,k)}^K)^T}{\sqrt{d_z}}, \]
RPE in Transformer-XL

\[ e_{ij} = \frac{(x_i W^Q + u)(x_j W^K)^T + (x_i W^Q - v)(s_{i-j} W^R)^T}{\sqrt{d_z}} \]
Huang’s RPE

\[ e_{ij} = \frac{(x_i W^Q + p_{ii})(x_j W^K + p_{jj})}{\sqrt{d_z} - p_{ii} p_{jj}} \]  
(11)
RPE in SASA

\[ e_{ij} = \frac{(x_i W^Q)(x_j W^K + \text{concat}(p_{\delta\tilde{x}}^K, p_{\delta\tilde{y}}^K))^T}{\sqrt{d_z}}, \quad (12) \]

\[ \delta\tilde{x} = \tilde{x}_i - \tilde{x}_j \]

\[ \delta\tilde{y} = \tilde{y}_i - \tilde{y}_j \]
RPE in Axial-Deeplab

\[ z_l = \sum_{j=1}^{n} \alpha_{ij} (x_j W^\nu + p^\nu_{ij}) \]

\[ e_{ij} = \frac{(x_i W^Q + p^Q_{ij})(x_j W^\kappa + p^\kappa_{ij})^T - p^Q_{ij} p^\kappa_{ij}}{\sqrt{d_z}} \]

\[ \alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{n} \exp(e_{ik})} , \]
New relative position encoding methods

• Bias Mode
• Contextual Mode
• A Piecewise Index Function
• 2D Relative Position Calculation
Bias and Contextual Mode

\[ e_{ij} = \frac{(x_i W^Q)(x_j W^K)^T + b_{ij}}{\sqrt{d_z}}, \quad (13) \]

\[ b_{ij} = r_{ij}, \quad (14) \]

\[ b_{ij} = (x_i W^Q)r_{ij}^T, \quad (15) \]

\[ b_{ij} = (x_i W^Q)(r_{ij}^K)^T + (x_j W^K)(r_{ij}^Q)^T, \quad (16) \]

\[ z_i = \sum_{j=1}^{n} \alpha_{ij} (x_j W^V + r_{ij}^V), \quad (17) \]
A Piecewise Index Function

- mapping a relative distance into an integer in finite set

\[
clip(x, k) = \max(-k, \min(k, x)), \quad (9)
\]

\[
g(x) = \begin{cases} 
[x], & |x| \leq \alpha \\
\text{sign}(x) \times \min(\beta, [\alpha + \frac{\ln(|x|/\alpha)}{\ln(\gamma/\alpha)} (\beta - \alpha)]), & |x| > \alpha 
\end{cases} \quad (18)
\]
2D Relative Position Calculation

- Euclidean method:
  \[ \begin{align*}
  r_{ij} &= P_{I(i,j)}, \\
  I(i, j) &= g(\sqrt{({\tilde{x}}_i - {\tilde{x}}_j)^2 + ({\tilde{y}}_i - {\tilde{y}}_j)^2}).
  \end{align*} \] (19) (20)

- Quantization method
  \[ I(i, j) = g(\text{quant}(\sqrt{({\tilde{x}}_i - {\tilde{x}}_j)^2 + ({\tilde{y}}_i - {\tilde{y}}_j)^2})). \] (21)
2D Relative Position Calculation

- Cross method
  \[ r_{ij} = p_{\tilde{x}(i,j)} + p_{\tilde{y}(i,j)} \]  \( (22) \)
  \[ I_{\tilde{x}}(i,j) = g(\tilde{x}_i - \tilde{x}_j) \]  \( (23) \)
  \[ I_{\tilde{y}}(i,j) = g(\tilde{y}_i - \tilde{y}_j) \]  \( (24) \)

- Product method
  \[ r_{ij} = p_{I_{\tilde{x}}(i,j), I_{\tilde{y}}(i,j)} \]  \( (25) \)
Experiments

- Compare each of the proposed methods
- Compare proposed methods to SOTA image classification and object detection
- Visualization of RPE
Implementation details

- DeiT with same training settings
- RPE added to keys
- Ratio for piecewise function constants
  - \( \alpha:\beta:\gamma = 1:2:8 \)
  - \[ g(x) = \begin{cases} 
  [x], & |x| \leq \alpha \\
  \text{sign}(x) \times \min(\beta, \lceil \alpha + \frac{\ln(|x|/\alpha)}{\ln(\gamma/\alpha)} (\beta - \alpha) \rceil), & |x| > \alpha 
  \end{cases} \]
  \hspace{1cm} (18)
- Add 1 to \( \beta \) (# buckets) for RPE of classification token
### Directed vs. Undirected Bias vs. Contextual

<table>
<thead>
<tr>
<th>Method based on DeiT-S [21]</th>
<th>Is Directed</th>
<th>Mode</th>
<th>Top-1 Acc(%)</th>
<th>Δ Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original [21]</td>
<td>-</td>
<td>-</td>
<td>79.9</td>
<td>-</td>
</tr>
<tr>
<td>Euclidean</td>
<td>×</td>
<td>bias</td>
<td>80.1</td>
<td>+0.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>contextual</td>
<td>80.4</td>
<td>+0.5</td>
</tr>
<tr>
<td>Quantization</td>
<td>×</td>
<td>bias</td>
<td>80.3</td>
<td>+0.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>contextual</td>
<td>80.5</td>
<td>+0.6</td>
</tr>
<tr>
<td>Cross</td>
<td>✓</td>
<td>bias</td>
<td>80.5</td>
<td>+0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>contextual</td>
<td>80.8</td>
<td>+0.9</td>
</tr>
<tr>
<td>Product</td>
<td>✓</td>
<td>bias</td>
<td>80.5</td>
<td>+0.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>contextual</td>
<td>80.9</td>
<td>+1.0</td>
</tr>
</tbody>
</table>
### Shared v.s. Unshared

<table>
<thead>
<tr>
<th>Mode</th>
<th>Shared</th>
<th>#Param. (M)</th>
<th>MACs (M)</th>
<th>Top-1 Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>✗</td>
<td>22.05</td>
<td>4613</td>
<td>80.54 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>22.05</td>
<td>4613</td>
<td>80.05 ± 0.04</td>
</tr>
<tr>
<td>Contextual</td>
<td>✗</td>
<td>22.28</td>
<td>4659</td>
<td>80.99 ± 0.16</td>
</tr>
<tr>
<td></td>
<td>✓</td>
<td>22.09</td>
<td>4659</td>
<td>80.89 ± 0.04</td>
</tr>
</tbody>
</table>
Piecewise v.s. Clip

Image Classification with DeiT

<table>
<thead>
<tr>
<th>Function</th>
<th>Mode</th>
<th>Top-1 Acc(%)</th>
<th>Top-5 Acc(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>clip</td>
<td>bias</td>
<td>80.1</td>
<td>94.9</td>
</tr>
<tr>
<td></td>
<td>contextual</td>
<td>80.9</td>
<td>95.5</td>
</tr>
<tr>
<td>piecewise</td>
<td>bias</td>
<td>80.0</td>
<td>95.0</td>
</tr>
<tr>
<td></td>
<td>contextual</td>
<td>80.9</td>
<td>95.5</td>
</tr>
</tbody>
</table>

Object Detection with DETR

| 5 | sinusoid | ctx clip | 9 x 9 | 150   | 40.4 | 60.9 | 42.4 | 19.1 | 43.7 | 59.8 |
| 6 | sinusoid | contextual| 9 x 9 | 150   | 40.8 | 61.5 | 42.5 | 18.5 | 44.4 | 60.5 |

\[
clip(x, k) = \max(-k, \min(k, x)),
\]

(9)

\[
g(x) = \begin{cases} 
|x|, & \text{if } |x| \leq \alpha \\
\text{sign}(x) \times \min(\beta, \frac{\ln(|x|/\alpha)}{\ln(\gamma/\alpha)} (\beta - \alpha)), & \text{if } |x| > \alpha 
\end{cases}
\]

(18)
Number of buckets

![Graph showing top-1 accuracy (%) vs number of buckets]

- Top-1 Accuracy (%)
  - 81.0
  - 80.5
  - 80.0

- Number of buckets
  - 0
  - 10
  - 26
  - 50
  - 82
  - 122
  - 226
  - 362
  - 730

- Key points:
  - 79.9
  - 80.7
  - 80.8
  - 80.9
  - 81.0
  - 80.8
  - 80.9
Component-wise analysis

<table>
<thead>
<tr>
<th>#</th>
<th>Abs Pos.</th>
<th>$p_{ij}^Q$</th>
<th>$p_{ij}^K$</th>
<th>$p_{ij}^Y$</th>
<th>Top-1</th>
<th>Top-5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>learnable</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>79.9</td>
<td>95.0</td>
</tr>
<tr>
<td>2</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>77.6 (-2.3)</td>
<td>93.8</td>
</tr>
<tr>
<td>3</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>80.9 (+1.0)</td>
<td>95.4</td>
</tr>
<tr>
<td>4</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>80.9 (+1.0)</td>
<td>95.3</td>
</tr>
<tr>
<td>5</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>80.2 (+0.3)</td>
<td>95.0</td>
</tr>
<tr>
<td>6</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>81.0 (+1.1)</td>
<td>95.5</td>
</tr>
<tr>
<td>7</td>
<td>×</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>81.3 (+1.4)</td>
<td><strong>95.7</strong></td>
</tr>
<tr>
<td>8</td>
<td>learnable</td>
<td>✓</td>
<td>×</td>
<td>×</td>
<td>80.9 (+1.0)</td>
<td>95.5</td>
</tr>
<tr>
<td>9</td>
<td>learnable</td>
<td>×</td>
<td>✓</td>
<td>×</td>
<td>80.9 (+1.0)</td>
<td>95.5</td>
</tr>
<tr>
<td>10</td>
<td>learnable</td>
<td>×</td>
<td>×</td>
<td>✓</td>
<td>80.2 (+0.3)</td>
<td>95.1</td>
</tr>
<tr>
<td>11</td>
<td>learnable</td>
<td>✓</td>
<td>✓</td>
<td>×</td>
<td>81.1 (+1.2)</td>
<td>95.4</td>
</tr>
<tr>
<td>12</td>
<td>learnable</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td><strong>81.4</strong> (+1.5)</td>
<td>95.6</td>
</tr>
</tbody>
</table>
Complexity Analysis

- **Inefficient Implementation:**
  \[ y_{ij} = (x_i \mathbf{W}) p_{I(i,j)}^T. \]

- **Efficient Implementation:**
  \[ z_{i,t} = (x_i \mathbf{w}) p_t^T, t \in \{I(i,j) | i, j \in [0, n]\}, \]
  \[ y_{ij} = z_{i,I(i,j)}. \]
### Comparison on image classification

<table>
<thead>
<tr>
<th>Model</th>
<th>#Param.</th>
<th>Input</th>
<th>MACs (M)</th>
<th>Top-1 Acc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ConvNets</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ResNet-50 [10]</td>
<td>25M</td>
<td>$224^2$</td>
<td>4121</td>
<td>79.0</td>
</tr>
<tr>
<td>RegNetY-4.0GF [14]</td>
<td>21M</td>
<td>$224^2$</td>
<td>4012</td>
<td>79.4</td>
</tr>
<tr>
<td>EfficientNet-B1 [20]</td>
<td>8M</td>
<td>$240^2$</td>
<td>712</td>
<td>79.1</td>
</tr>
<tr>
<td>EfficientNet-B5 [20]</td>
<td>30M</td>
<td>$456^2$</td>
<td>10392</td>
<td>83.6</td>
</tr>
<tr>
<td><strong>Transformers</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-B/16 [7]</td>
<td>86M</td>
<td>$384^2$</td>
<td>55630</td>
<td>77.9</td>
</tr>
<tr>
<td>ViT-L/16 [7]</td>
<td>307M</td>
<td>$384^2$</td>
<td>191452</td>
<td>76.5</td>
</tr>
<tr>
<td>DeiT-Ti [21]</td>
<td>5M</td>
<td>$224^2$</td>
<td>1261</td>
<td>72.2</td>
</tr>
<tr>
<td>CPVT-Ti(0-5) [13]</td>
<td>6M</td>
<td>$224^2$</td>
<td>1262</td>
<td>73.4</td>
</tr>
<tr>
<td><strong>DeiT-Ti with iRPE-K (Ours)</strong></td>
<td>6M</td>
<td>$224^2$</td>
<td>1284</td>
<td>73.7</td>
</tr>
<tr>
<td>DeiT-S [21]</td>
<td>22M</td>
<td>$224^2$</td>
<td>4613</td>
<td>79.9</td>
</tr>
<tr>
<td>CPVT-S(0-5) [3]</td>
<td>23M</td>
<td>$224^2$</td>
<td>4616</td>
<td>80.5</td>
</tr>
<tr>
<td>DeiT-S(Shaw's) [17, 21]$^+$</td>
<td>22M</td>
<td>$224^2$</td>
<td>4659</td>
<td>80.9</td>
</tr>
<tr>
<td>DeiT-S(trans.-XL's) [4, 21]$^+$</td>
<td>23M</td>
<td>$224^2$</td>
<td>4828</td>
<td>80.8</td>
</tr>
<tr>
<td>DeiT-S(Huang's) [11, 21]$^+$</td>
<td>22M</td>
<td>$224^2$</td>
<td>4706</td>
<td>81.0</td>
</tr>
<tr>
<td>DeiT-S(SASA's) [16, 21]$^+$</td>
<td>22M</td>
<td>$224^2$</td>
<td>4639</td>
<td>80.8</td>
</tr>
<tr>
<td><strong>DeiT-S with iRPE-K (Ours)</strong></td>
<td>22M</td>
<td>$224^2$</td>
<td>4659</td>
<td>80.9</td>
</tr>
<tr>
<td>DeiT-S with iRPE-QK (Ours)</td>
<td>22M</td>
<td>$224^2$</td>
<td>4706</td>
<td>81.1</td>
</tr>
<tr>
<td>DeiT-S with iRPE-QKV (Ours)</td>
<td>22M</td>
<td>$224^2$</td>
<td>4885</td>
<td>81.4</td>
</tr>
<tr>
<td><strong>DeiT-B [21]</strong></td>
<td>86M</td>
<td>$224^2$</td>
<td>17592</td>
<td>81.8</td>
</tr>
<tr>
<td>CPVT-B(0-5) [13]</td>
<td>86M</td>
<td>$224^2$</td>
<td>17508</td>
<td>83.9</td>
</tr>
<tr>
<td><strong>DeiT-B with iRPE-K (Ours)</strong></td>
<td>87M</td>
<td>$224^2$</td>
<td>17684</td>
<td>82.4</td>
</tr>
</tbody>
</table>
Comparison on object detection

<table>
<thead>
<tr>
<th>#</th>
<th>Abs Pos.</th>
<th>Rel Pos.</th>
<th>#buckets</th>
<th>epoch</th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_{S}$</th>
<th>AP$_{M}$</th>
<th>AP$_{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sinusoid</td>
<td>none</td>
<td>-</td>
<td>150</td>
<td>39.5</td>
<td>60.3</td>
<td>41.4</td>
<td>17.5</td>
<td>43.0</td>
<td>59.1</td>
</tr>
<tr>
<td>2</td>
<td>none</td>
<td>none</td>
<td>-</td>
<td>150</td>
<td>30.4</td>
<td>52.5</td>
<td>30.2</td>
<td>9.4</td>
<td>31.2</td>
<td>50.5</td>
</tr>
<tr>
<td>3</td>
<td>sinusoid</td>
<td>bias</td>
<td>9 × 9</td>
<td>150</td>
<td>40.6</td>
<td>61.2</td>
<td>42.8</td>
<td>19.0</td>
<td>43.9</td>
<td>60.2</td>
</tr>
<tr>
<td>4</td>
<td>none</td>
<td>contextual</td>
<td>9 × 9</td>
<td>150</td>
<td>38.7</td>
<td>60.1</td>
<td>40.4</td>
<td>18.2</td>
<td>41.8</td>
<td>56.7</td>
</tr>
<tr>
<td>5</td>
<td>sinusoid</td>
<td>ctx clip</td>
<td>9 × 9</td>
<td>150</td>
<td>40.4</td>
<td>60.9</td>
<td>42.4</td>
<td>19.1</td>
<td>43.7</td>
<td>59.8</td>
</tr>
<tr>
<td>6</td>
<td>sinusoid</td>
<td>contextual</td>
<td>9 × 9</td>
<td>150</td>
<td>40.8</td>
<td>61.5</td>
<td>42.5</td>
<td>18.5</td>
<td>44.4</td>
<td>60.5</td>
</tr>
<tr>
<td>7</td>
<td>sinusoid</td>
<td>contextual</td>
<td>15 × 15</td>
<td>150</td>
<td>40.8</td>
<td>61.7</td>
<td>42.6</td>
<td>18.5</td>
<td>44.2</td>
<td>61.2</td>
</tr>
<tr>
<td>8</td>
<td>sinusoid</td>
<td>none</td>
<td>-</td>
<td>300</td>
<td>40.6</td>
<td>61.6</td>
<td>-</td>
<td>19.9</td>
<td>44.3</td>
<td>60.2</td>
</tr>
<tr>
<td>9</td>
<td>sinusoid</td>
<td>contextual</td>
<td>9 × 9</td>
<td>300</td>
<td>42.3</td>
<td>62.8</td>
<td>44.3</td>
<td>20.7</td>
<td>46.2</td>
<td>61.1</td>
</tr>
</tbody>
</table>
Visualization

- Visualize the extra for different positions.
- RPE makes patches focus more in block 0.
Conclusion

- Review existing relative position encoding methods
- Propose four methods dedicated to visual transformers.
- The experiments show effectiveness of their methods on both classification and detection tasks.
- Their methods are easily adapted to other models.
Reference