Faster Transformers for Document Summarization
Zaid Nabulsi, Dian Ang Yap, Vineet Kosaraju

Document summarization has been done through vanilla RNNs, RL agents, and transformers. Transformers are very promising but are difficult to train as there attention layers serve as bottleneck. We present architectural design modifications to improve both efficiency and performance.

**Task: Long Document Summarization**

**Overall Architecture**

- **Encoder**
  - Linear (Gather)
  - Add and Norm
  - Strided Split
  - Core Attention
  - Strided Split
  - Multi-Head Module
  - Multi-Head Module
  - Multi-Head Module
  - Linear
  - Linear
  - Linear
- **Decoder**
  - Linear
  - Linear
  - Linear
  - Linear
  - Add and Norm
  - Feed Forward

**Input**

- **Output**

**Approach & Methods**

- **Transformer Architecture**
  - Encoder
  - Decoder
  - Core Attention
  - Add and Norm
  - Fixed Forward
  - Add and Norm

**Strided Neighborhood Attention**

- **A**
  - Mask
  - Multi-Head Module
  - Multi-Head Module
  - Multi-Head Module
  - Linear
  - Linear
  - Linear
  - Linear
  - Linear
  - Linear
  - Linear
  - Linear
  - Linear
  - Linear

**Results**

- **Attention**
  - Accuracy
  - Perplexity
  - Speed (Tokens/s)
  - Theoretical Runtime

<table>
<thead>
<tr>
<th>Attention</th>
<th>Training</th>
<th>Validation</th>
<th>Training</th>
<th>Validation</th>
<th>Speed</th>
<th>Theoretical Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>56.12</td>
<td>56.55</td>
<td>7.65</td>
<td>9.26</td>
<td>5.96</td>
<td>54.48</td>
</tr>
<tr>
<td>Conv.</td>
<td>56.51</td>
<td>56.52</td>
<td>7.50</td>
<td>9.41</td>
<td>6.28</td>
<td>57.70</td>
</tr>
<tr>
<td>Strided</td>
<td>56.62</td>
<td>56.77</td>
<td>7.49</td>
<td>9.01</td>
<td>6.29</td>
<td>58.18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Baseline Attention</th>
<th>Convolutional Attention</th>
<th>Strided Attention</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROUGE-1 Recall</td>
<td>38.60</td>
<td>39.26</td>
</tr>
<tr>
<td>ROUGE-1 Precision</td>
<td>41.90</td>
<td>41.19</td>
</tr>
<tr>
<td>ROUGE-1 F Score</td>
<td>38.82</td>
<td>38.77</td>
</tr>
<tr>
<td>ROUGE-2 Recall</td>
<td>16.64</td>
<td>16.86</td>
</tr>
<tr>
<td>ROUGE-2 Precision</td>
<td>18.47</td>
<td>18.06</td>
</tr>
<tr>
<td>ROUGE-2 F Score</td>
<td>16.89</td>
<td>16.80</td>
</tr>
<tr>
<td>ROUGE-L Recall</td>
<td>35.62</td>
<td>36.38</td>
</tr>
<tr>
<td>ROUGE-L Precision</td>
<td>38.76</td>
<td>38.27</td>
</tr>
<tr>
<td>ROUGE-L F Score</td>
<td>35.87</td>
<td>35.97</td>
</tr>
</tbody>
</table>

**Conclusions**

- Presented two novel models with architectural improvements to transformers that allow for more efficient training while maintaining (and even exceeding) comparable metrics to existing state-of-the-art methods on document summarization.
- As next steps, combining the models might result in even better performance.

**References**


We would like to acknowledge the help of Kevin Clark and Abigail See in mentoring us on this project and helping us understand theoretical contributions, as well as giving feedback on datasets, experiments, visualizations, and our milestone reports.