Long Document Summarization

Qi Zeng
qizeng2@illinois.edu
Overview: Long-doc Summarization

● Challenges
● Methods
  ○ X-formers: Efficient Transformers: A Survey
  ○ Evaluation: Long Range Arena: A Benchmark for Efficient Transformers
  ○ HEPOS: Efficient Attentions for Long Document Summarization
  ○ Skyformer: Remodel Self-Attention with Gaussian Kernel and Nystr"om Method
  ○ Skeinformer: Sketching as a Tool for Understanding and Accelerating Self-attention for Long Sequences
● Resources
  ○ Datasets and tools
Introduction
Quick Recap: Short-doc Summarization

Source Document

( @entity0 ) wanted: film director, must be eager to shoot footage of golden lassos and invisible jets. <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie ( the hollywood reporter first broke the story ). <eos> @entity5 was announced as director of the movie in november. <eos> @entity0 obtained a statement from @entity13 that says, " given creative differences, @entity13 and @entity5 have decided not to move forward with plans to develop and direct ' @entity9 ' together . <eos> " ( @entity0 and @entity13 are both owned by @entity16 . <eos> ) the movie, starring @entity18 in the title role of the @entity21 princess, is still set for release on june 00 , 0000. <eos> it 's the first theatrical movie centering around the most popular female superhero. <eos> @entity18 will appear beforehand in " @entity25 v. @entity26 : @entity27 , " due out march 00 , 0000. <eos> in the meantime, @entity13 will need to find someone new for the director 's chair. <eos>

Ground truth Summary

@entity5 is no longer set to direct the first " @entity9 " theatrical movie <eos> @entity5 left the project over " creative differences " <eos> movie is currently set for 0000

words-lvt2k

@entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie <eos> @entity13 and @entity5 have decided not to move forward with plans to develop <eos> @entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie

words-lvt2k-hieratt

@entity5 is leaving the upcoming " @entity9 " movie <eos> the movie is still set for release on june 00 , 0000 <eos> @entity5 is still set for release on june 00 , 0000

words-lvt2k-temp-att

@entity0 confirms that @entity5 is leaving the upcoming " @entity9 " movie <eos> the movie is the first film to around the most popular female actor <eos> @entity18 will appear in " @entity25 , " due out march 00 , 0000

Figure 1: Example output of the attention-based summarization (ABS) system. The heatmap represents a soft alignment between the input (right) and the generated summary (top). The columns represent the distribution over the input after generating each word.
Challenges: From short-doc to long-doc

- General challenges for long sequence processing
  - Computation complexity due to length of sequence
  - The inter-dependency among processing units
  - The difficulty of information aggregation across long context
  - The lack of available datasets due to costly annotation

- Task-Specific challenges
  - **Summarization**: encoder-decoder attentions collaborate and dynamically pinpoint salient content in the source as the summary is decoded (Huang et al, NAACL 2021)
  - **Machine Translation**: existing approaches simply introduce the representations of context sentences without explicitly characterizing the inter-sentence reasoning process; and feed ground-truth target contexts as extra inputs at the training time, thus facing the problem of exposure bias (Zhang et al, NAACL 2021)
  - **Document Ranking**: remote connections between terms are ignored or captured by simple patterns (Zhou et al, 2021)
  - **Event Argument Extraction**: Sentence-level argument extraction approaches cannot handle the cross-sentence trigger-argument distribution and the existence of multiple events within one document (our paper in submission)
Methods
Method: Long-Doc Summarization

- Retrieval
- Graph
- Memory
- Hierarchy
- Attention
X-former: Transformer Acceleration Methods

- **Pruning redundant attention head**


- **Model size reduction with knowledge distillation**


- **Attention Approximation/Acceleration**


---

X-former: fast attention for efficient transformers
Fixed Patterns (FP)

- Limit the field of view to fixed, predefined patterns

Blockwise Pattern: BlockBERT [8]

Strided/Dilated Pattern: Sparse Transformer [9]
Combination of Patterns (CP)

- Combine two or more distinct access patterns

(a) Random attention  (b) Window attention  (c) Global Attention  (d) BIGBIRD

BIGBIRD [12]
Learnable Patterns (LP)

- Learn the access pattern in a data-driven fashion
- Determine the token relevance and assign tokens to buckets or clusters

Routing Transformers [14]
Memory (M)

- Access multiple tokens at once with Global Memory
Low-Rank Methods (LR)

- Assume low-rank structure in the $N \times N$ matrix

Linformer [16]
X-former: Longformer

- Combines a local windowed attention with a task motivated global attention
  - Motivation of global attention: the optimal input representation varies by task

(a) Full $n^2$ attention  (b) Sliding window attention  (c) Dilated sliding window  (d) Global+sliding window
X-former: Performer

- **FAVOR+**: Fast Attention Via positive Orthogonal Random features
  - FA\[ A = \exp(QK^T) \approx \phi(Q)\phi(K)^T = Q'(K')^T \]
  - R+\[ \phi(x) = \frac{h(x)}{\sqrt{m}} (f_1(\omega_1^T x), ..., f_1(\omega_m^T x), ..., f_l(\omega_1^T x), ..., f_l(\omega_m^T x)) \]
  - O entangle different random samples $\omega_1, ..., \omega_m$ to be exactly orthogonal
X-former: Informer

- ProbSparse self-attention
- Self-attention distillation
X-former: Nyströmformer

- Nyström Method is widely adopted for matrix approximation
- Sample a subset of columns and rows
  - Rewrite attention matrix, approximate $S$ via the basic quadrature technique of the Nyström Method
  - Challenge
  - Using landmarks

$$ S = \text{softmax} \left( \frac{QK^T}{\sqrt{d_q}} \right) = \begin{bmatrix} A_S & B_S \\ F_S & C_S \end{bmatrix} \quad \tilde{S} = \begin{bmatrix} A_S \\ F_S A_S^+ B_S \end{bmatrix} = \begin{bmatrix} A_S \\ F_S \end{bmatrix} A_S^+ \begin{bmatrix} A_S & B_S \end{bmatrix} \quad \hat{S} = \left[ \text{softmax} \left( \frac{QK^T}{\sqrt{d_q}} \right) \right]_{n \times m} \begin{bmatrix} A_S^+ & \text{softmax} \left( \frac{QK^T}{\sqrt{d_q}} \right) \end{bmatrix}_{m \times n} $$

softmax

Nyström approximation
Long Range Arena: A Benchmark for Efficient Transformers

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, Donald Metzler
This paper proposes a systematic and unified benchmark specifically focused on evaluating model quality under long-context scenarios.
LRA: Evaluation Benchmark for X-formers

Six classification tasks for transformer encoder

- **ListOps**: seq_len=2k
  - INPUT: \([\text{MAX } 4 \ 3 \ \text{MIN } 2 \ 3 \ ] \ 1 \ 0 \ [\text{MEDIAN } 1 \ 5 \ 8 \ 9, \ 2]]\)
  - OUTPUT: 5

- **Byte-level Text Classification**: IMDb reviews, seq_len=4k

- **Byte-level Document Retrieval**: ANN dataset, seq_len=4k, two-tower setting

- **Image Classification**: CIFAR-10, seq_len=1k

- **Pathfinder** (long-range spatial dependency): seq_len=1k (32*32)

- **Pathfinder-X**: seq_len=16k (128*128)
LRA: Comparisons are made with 2-layer transformers

Why not larger or deeper transformers?

- Training a deep transformer from scratch requires large computational resources and much more data to converge, and therefore is not adopted by previous work.
- A shallow transformer structure, on the other hand, has been justified by previous work to be enough for fair comparison in attention acceleration performance.
- Pretrained models, like BERT, are trained for token-level text-based tasks, and are not suitable for image pixel sequences (as in Pathfinder and Image Classification), character sequences (as in Text Classification) and math operation sequences (as in ListOps).
Efficient Attentions for Long Document Summarization

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, Lu Wang
NAACL 2021
HEPOS: Efficient Attentions for Long Document Summarization
Sketching as a Tool for Understanding and Accelerating Self-attention for Long Sequences

Yifan Chen*, Qi Zeng*, Dilek Hakkani-Tur, Di Jin, Heng Ji, Yun Yang
Skeinformer: Background

Revisit Self-Attention

Figure 1: Architecture of the standard Transformer (Vaswani et al., 2017)
Skeinforme: Motivation - Sketching Framework

Linformer and Informer can be unified into the sketching framework:

- Sketching methods replace the original matrix B with its random sketch BS.
- Informer selects d important rows of $D^{-1}A$, which can be related to a sketched approximation $D^{-1}SS^TA$.
- Linformer adds two linear projection layers, $(QK^*/\sqrt{d})S \cdot S^TV$, which can be related to a sketched approximation $D^{-1}ASS^TV$. 


Skeinformer: Method

- **Column Sampling in the sketching form**
  - Sampling probability in sub-sampling matrix $S$
  - Advantages:
    - This sketching form circumvents the computation burden of Gaussian sketching
    - This sketching form allows the incorporation of information from $V$
    - The construction of $S$ theoretically guarantees the approximation performance in terms of Frobenius norm loss
  
  $p_i \propto \|(D^{-1}A)^{(i)}\|_2 \|V^{(i)}\|_2, \quad i = 1, 2, \ldots, n.$

- **Adaptive Row Normalization**
  - Filling the unselected columns with the averaged selected columns
  - Advantage: It allows the whole value matrix $V$ to participate in the computation, which will improve the efficiency of updating $W_V$

- **Pilot Sampling Reutilization**
  - Reproduce the d ros in the original self-attention output with an additional product
Algorithm 1: Skeinformer.

**Input:** query matrix $Q$, key matrix $K$, value matrix $V$ (all are $n$-by-$p$), and sub-sample size $d$

**Output:** Attention output matrix $R$ with the same shape as $V$

1. Uniformly sample $d$ indices $j_1, \cdots, j_d$ with replacement;
2. Construct the $d \times p$ matrix $Q_J$ as to the index set $J := \{j_k\}_{k=1}^d$, whose $k$-th row is $Q_{(j_k)}$;
3. Compute the matrix $B_J = \text{softmax} \left( Q_J K^T / \sqrt{p} \right)$; \hspace{2cm} // pilot sampling
4. Based on $B_J$, give the estimated sub-sampling probabilities $\{\hat{p}_i\}_{i=1}^n$ as in Equation (5);
5. With $\{\hat{p}_i\}_{i=1}^n$ sample $d$ indices $j'_1, \cdots, j'_d$ without replacement;
6. Construct the $d$-by-$p$ matrix $K_{J'}$ (resp., $V_{J'}$) according to the indices list $J' := \{j'_k\}_{k=1}^d$, whose $k$-th row is $K_{(j'_k)}$ (resp., $V_{(j'_k)}$);
7. Compute the two matrices $A'_{J'} = \exp \left( Q K_{J'}^T / \sqrt{p} \right)$, and $R_{J'} = A'_{J'} V_{J'}$; \hspace{2cm} // column sampling
8. Construct a length $n$ column vector $g$ whose $i$-th element is $(\prod_{k=1}^d a_{i j'_k})^{1/d}$, $\forall i \in [n]$;
9. Compute the row sum vector $d := A'_{J'} 1_d + (n - d)g$; \hspace{2cm} // adaptive row normalization
10. Denote the un-selected part of $V$ as $V_{(J')^C}$, and compute the vector $v = V_{(J')^C}^T 1_{n-d}$;
11. Obtain the intermediate output $R = \text{diag}(d^{-1})(R_{J'} + g v^T)$, where $d^{-1}$ is the element-wise inverse of $d$;
12. Compute $B_J V$ and assign it to the corresponding rows of $R$; \hspace{2cm} // pilot sampling reutilization
13. Return the matrix $R$ as the ultimate output of this algorithm;
## Skeinformer: Classification accuracy on LRA

<table>
<thead>
<tr>
<th>Models</th>
<th>Text</th>
<th>ListOps</th>
<th>Retrieval</th>
<th>Pathfinder</th>
<th>Image</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard (Vaswani et al. 2017)</td>
<td>57.69</td>
<td>38.15</td>
<td>80.10</td>
<td>73.59</td>
<td>37.97</td>
<td>57.50</td>
</tr>
<tr>
<td>· w/o dropout</td>
<td>59.44</td>
<td>38.17</td>
<td>79.35</td>
<td>72.35</td>
<td>37.58</td>
<td>57.38</td>
</tr>
<tr>
<td>V-mean</td>
<td>65.29</td>
<td>28.78</td>
<td>80.49</td>
<td>61.01</td>
<td>34.33</td>
<td>53.98</td>
</tr>
<tr>
<td>BigBird (Zaheer et al. 2020)</td>
<td>61.91</td>
<td>38.86</td>
<td>79.73</td>
<td>71.75</td>
<td>35.00</td>
<td>57.45</td>
</tr>
<tr>
<td>Performer (Choromanski et al. 2020)</td>
<td>57.67</td>
<td>37.70</td>
<td>75.69</td>
<td>56.50</td>
<td>37.40</td>
<td>52.99</td>
</tr>
<tr>
<td>Nystromformer (Xiong et al. 2021)</td>
<td>60.91</td>
<td>37.76</td>
<td>79.87</td>
<td>72.53</td>
<td>31.93</td>
<td>56.60</td>
</tr>
<tr>
<td>Reformer (Kitaev, Kaiser, and Levskaya 2020)</td>
<td>62.69</td>
<td>37.94</td>
<td>78.85</td>
<td>69.21</td>
<td>36.42</td>
<td>57.02</td>
</tr>
<tr>
<td>Linformer (Wang et al. 2020a)</td>
<td>58.52</td>
<td>37.97</td>
<td>77.40</td>
<td>55.57</td>
<td>37.48</td>
<td>53.39</td>
</tr>
<tr>
<td>· w/ unreduced JLT</td>
<td>59.12</td>
<td>37.48</td>
<td>79.39</td>
<td>68.45</td>
<td>35.96</td>
<td>56.08</td>
</tr>
<tr>
<td>Informer (Zhou et al. 2020)</td>
<td>61.55</td>
<td>38.43</td>
<td>80.88</td>
<td>59.34</td>
<td>36.55</td>
<td>55.35</td>
</tr>
<tr>
<td>· w/ padding mask</td>
<td>60.98</td>
<td>37.26</td>
<td>79.92</td>
<td>62.51</td>
<td>37.19</td>
<td>55.57</td>
</tr>
<tr>
<td><strong>Skeinformer</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>58.08</strong></td>
</tr>
<tr>
<td>· w/ uniform sampling</td>
<td>64.48</td>
<td>30.02</td>
<td>80.57</td>
<td>64.35</td>
<td>36.97</td>
<td>55.28</td>
</tr>
<tr>
<td>· w/o row normalization</td>
<td>60.67</td>
<td>37.69</td>
<td>78.67</td>
<td>66.35</td>
<td>37.06</td>
<td>56.09</td>
</tr>
<tr>
<td>· w/ simple row normalization</td>
<td>60.26</td>
<td>38.35</td>
<td>78.97</td>
<td>65.41</td>
<td>39.72</td>
<td>56.54</td>
</tr>
<tr>
<td>· w/o pilot sampling reutilization</td>
<td>62.39</td>
<td>38.12</td>
<td>79.88</td>
<td>71.53</td>
<td>37.20</td>
<td>57.83</td>
</tr>
</tbody>
</table>
### Skeinformer: Time and Space on LRA

<table>
<thead>
<tr>
<th>Models</th>
<th>step</th>
<th>Text time</th>
<th>accu</th>
<th>ListOps step</th>
<th>time</th>
<th>accu</th>
<th>Retrieval step</th>
<th>time</th>
<th>accu</th>
<th>Pathfinder step</th>
<th>time</th>
<th>accu</th>
<th>Image step</th>
<th>time</th>
<th>accu</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>11</td>
<td>51</td>
<td>8</td>
<td>9</td>
<td>22</td>
<td>4</td>
<td>20</td>
<td>53</td>
<td>4</td>
<td>21</td>
<td>14</td>
<td>4</td>
<td>4</td>
<td>21</td>
<td>4</td>
</tr>
<tr>
<td>w/o dropout</td>
<td>6</td>
<td>39</td>
<td>16</td>
<td>6</td>
<td>20</td>
<td>8</td>
<td>16</td>
<td>42</td>
<td>8</td>
<td>18</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>V-mean</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>4</td>
<td>1</td>
<td>27</td>
<td>4</td>
<td>1</td>
<td>16</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>BigBird</td>
<td>6</td>
<td>21</td>
<td>2</td>
<td>9</td>
<td>17</td>
<td>2</td>
<td>15</td>
<td>22</td>
<td>2</td>
<td>17</td>
<td>18</td>
<td>2</td>
<td>3</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td>Performer</td>
<td>30</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td>9</td>
<td>10</td>
<td>2</td>
<td>5</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Nystromformer</td>
<td>12</td>
<td>12</td>
<td>2</td>
<td>10</td>
<td>12</td>
<td>2</td>
<td>17</td>
<td>13</td>
<td>2</td>
<td>23</td>
<td>20</td>
<td>4</td>
<td>3</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>Reformer</td>
<td>7</td>
<td>11</td>
<td>2</td>
<td>14</td>
<td>8</td>
<td>2</td>
<td>26</td>
<td>11</td>
<td>1</td>
<td>24</td>
<td>9</td>
<td>2</td>
<td>5</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>Linformer</td>
<td>7</td>
<td>8</td>
<td>2</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>23</td>
<td>8</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>2</td>
<td>4</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>w/ unreduced JLT</td>
<td>6</td>
<td>37</td>
<td>16</td>
<td>4</td>
<td>21</td>
<td>8</td>
<td>24</td>
<td>36</td>
<td>16</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td>4</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Informer</td>
<td>8</td>
<td>33</td>
<td>8</td>
<td>10</td>
<td>22</td>
<td>8</td>
<td>25</td>
<td>37</td>
<td>4</td>
<td>9</td>
<td>26</td>
<td>8</td>
<td>7</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td>w/ mask</td>
<td>7</td>
<td>26</td>
<td>4</td>
<td>4</td>
<td>22</td>
<td>4</td>
<td>31</td>
<td>36</td>
<td>2</td>
<td>7</td>
<td>16</td>
<td>4</td>
<td>5</td>
<td>23</td>
<td>2</td>
</tr>
<tr>
<td>Skeinformer</td>
<td>6</td>
<td>10</td>
<td>2</td>
<td>7</td>
<td>10</td>
<td>2</td>
<td>20</td>
<td>11</td>
<td>1</td>
<td>18</td>
<td>9</td>
<td>2</td>
<td>6</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>w/ US</td>
<td>8</td>
<td>8</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>1</td>
<td>26</td>
<td>7</td>
<td>1</td>
<td>19</td>
<td>7</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>w/ RN</td>
<td>6</td>
<td>25</td>
<td>8</td>
<td>6</td>
<td>16</td>
<td>4</td>
<td>14</td>
<td>56</td>
<td>16</td>
<td>13</td>
<td>11</td>
<td>2</td>
<td>5</td>
<td>16</td>
<td>2</td>
</tr>
<tr>
<td>w/ SRN</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>19</td>
<td>8</td>
<td>1</td>
<td>14</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>w/o PSR</td>
<td>7</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>7</td>
<td>1</td>
<td>22</td>
<td>9</td>
<td>1</td>
<td>17</td>
<td>7</td>
<td>1</td>
<td>6</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>
Skeinformer: Experiment conclusions

- **Space Efficiency**: Skeinformer requires less space and enables larger batch size.
- **Time Efficiency**: Skeinformer has consistently less time consumption.
- **Fast Convergence**: Skeinformer efficiently converges to the long-time limit.
- **Comparable General Performance**: Most O(n) attention acceleration methods have comparable performance with vanilla attention.
Skyformer: Remodel Self-Attention with Gaussian Kernel and Nyström Method

Yifan Chen*, Qi Zeng*, Heng Ji, Yun Yang
NeurIPS 2021
Skyformer: Background

Transformers are expensive to train

- **Quadratic Complexity**
  - Quadratic time and space complexity in the self-attention mechanism
  - Transformers cannot support long sequence processing and large batch size

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{p}} \right) V := D^{-1} AV
\]

- **Training Instability**
  - Small perturbations in parameter updates tend to be amplified, resulting in significant disturbances in the model output
  - Training process is sensitive to hyper-parameters
Skyformer: Background (cont’)

Kernel methods may be the answer to both challenges (quadratic complexity and training instability)

- Connections between Self-attention and Gaussian Kernels:
  - The un-normalized attention score matrix can be formed via basic matrix operations on an empirical Gaussian Kernel matrix
  - The form of Gaussian kernels has the natural interpretation of assigning “attention” to different tokens
  - Gaussian kernels automatically perform the normalization similar to how softmax does
Skyformer: Method - Kernelized Attention

- Kernelized Attention replaces the softmax structure with a Gaussian kernel

\[
\text{Kernelized-Attention}(Q, K, V) = CV := \kappa \left( \frac{Q}{p^{1/4}}, \frac{K}{p^{1/4}} \right) V
\]

- It can be rewritten in terms of the un-normalized attention score matrix as

\[
\text{Kernelized-Attention}(Q, K, V) = D_Q^{-1/2} \cdot A \cdot D_K^{-1/2}
\]
Skyformer: a modified Nyström method

Complete the matrix into a PSD matrix $\tilde{B}$

$$\tilde{B} := \phi \left( \begin{pmatrix} Q \\ K \end{pmatrix}, \begin{pmatrix} Q \\ K \end{pmatrix} \right), \quad (B =: (I, 0)\tilde{B}(0, I)^T)$$

Approximate $\tilde{B}$ with $\tilde{\tilde{B}}$ through

$$\tilde{\tilde{B}} = \tilde{B}S(S^T\tilde{B}S)^\dagger S^T\tilde{B}$$

The final approximation will be given as

$$\tilde{\tilde{B}} := (I, 0)\tilde{B}(0, I)^T$$
Skyformer: Approximation evaluation results

[Graphs showing the spectral norm loss for different initializations and pretrainings with varying numbers of features.]
# Skyformer: Classification accuracy on LRA

<table>
<thead>
<tr>
<th>Model</th>
<th>Text</th>
<th>ListOps</th>
<th>Retrieval</th>
<th>Pathfinder</th>
<th>Image</th>
<th>AVG.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Attention</td>
<td>61.95</td>
<td>38.37</td>
<td>80.69</td>
<td>65.26</td>
<td>40.57</td>
<td>57.37</td>
</tr>
<tr>
<td>Kernelized Attention</td>
<td>60.22</td>
<td>38.78</td>
<td>81.77</td>
<td>70.73</td>
<td>41.29</td>
<td>58.56</td>
</tr>
<tr>
<td>Nystromformer</td>
<td>64.83</td>
<td>38.51</td>
<td>80.52</td>
<td>69.48</td>
<td>41.30</td>
<td>58.93</td>
</tr>
<tr>
<td>Linformer</td>
<td>58.93</td>
<td>37.45</td>
<td>78.19</td>
<td>60.93</td>
<td>37.96</td>
<td>54.69</td>
</tr>
<tr>
<td>Informer</td>
<td>62.64</td>
<td>32.53</td>
<td>77.57</td>
<td>57.83</td>
<td>38.10</td>
<td>53.73</td>
</tr>
<tr>
<td>Performer</td>
<td>64.19</td>
<td>38.02</td>
<td>80.04</td>
<td>66.30</td>
<td>41.43</td>
<td>58.00</td>
</tr>
<tr>
<td>Reformer</td>
<td>62.93</td>
<td>37.68</td>
<td>78.99</td>
<td>66.49</td>
<td>48.87</td>
<td>58.99</td>
</tr>
<tr>
<td>BigBird</td>
<td>63.86</td>
<td>39.25</td>
<td>80.28</td>
<td>68.72</td>
<td>43.16</td>
<td>59.05</td>
</tr>
<tr>
<td><strong>Skyformer</strong></td>
<td>64.70</td>
<td>38.69</td>
<td>82.06</td>
<td>70.73</td>
<td>40.77</td>
<td><strong>59.39</strong></td>
</tr>
</tbody>
</table>
Skyformer: Conclusion

- We revisit the intrinsic connection between self-attention and kernel methods, and explore a new kernel-based structure to stabilize the training of Transformers
- We approximate the Kernelized Attention via low dimensional randomized sketches by adapting the Nystrom method to a non-PSD matrix
- We conduct extensive experiments showing that Skyformer achieves comparable performance to the original self-attention with fewer computational costs
Resources
Resource: Project Topic

- Knowledge-enhanced Abstractive Summarization
- Possible extensions
  - Adaptive (plug-in) layers for Knowledge Infusion
  - A long transformer specializing in handling masks
  - Concept Representation learning
  - IE-assisted consistency checking
  - Evaluation benchmark
Resources: Datasets and Tools

- Long summarization datasets
  - GovReport
    https://drive.google.com/drive/folders/128KyqPTwZ0Si9RV_lX-md2dcHeRTUHkr
  - ArXiv
    https://drive.google.com/file/d/1b3rmCSIoh6VhD4HKWjl4HOW-cSwcwbeC/view?usp=sharing
  - PubMed
    https://drive.google.com/file/d/1lsqsqFi3W-pE1SqNZI0s8NR9rC1tsja/view?usp=sharing

- BART for quick experiment setup
  - https://huggingface.co/docs/transformers/model_doc/bart
Thanks!