Big Bird: Transformers for Longer Sequences

Manzil Zaheer MANZILZ@GOOGLE.COM
Guru Guruganesh GURUG@GOOGLE.COM
Google Research,
Mountain View, CA, USA

Avinava Dubey AVINAVADUBEY@GOOGLE.COM
Joshua Ainslie, Chris Alberti, Santiago Ontanon,
Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang,
Amr Ahmed
Google Research, USA

Intro & Overview

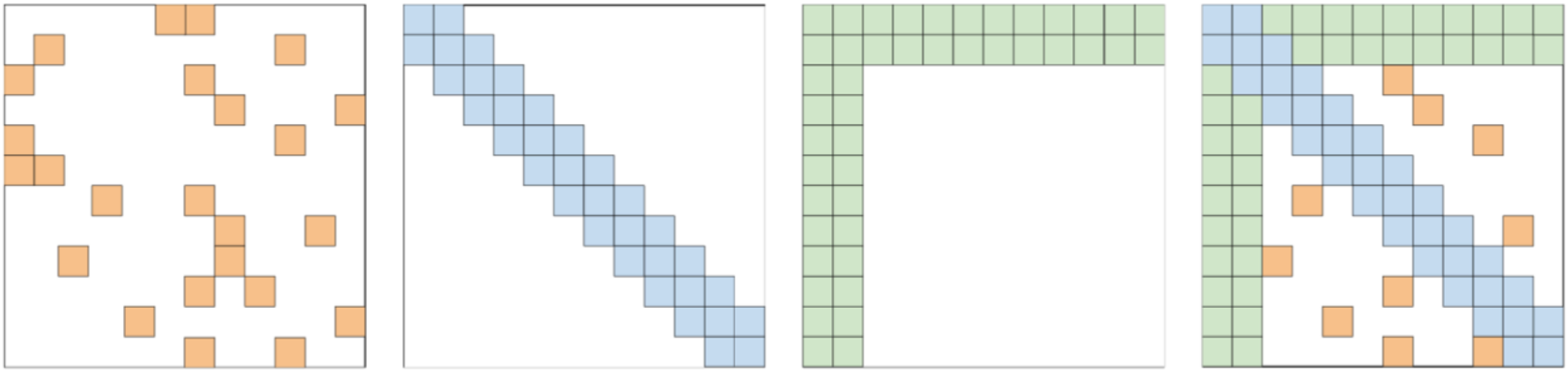
Limitations of Previous Transformers-Based Models

$O(n^2)$ quadratic dependency (mainly in terms of memory)

To remedy this

- BigBird, a sparse attention mechanism that reduces this quadratic dependency to linear $O(n)$
- theoretical analysis
- question answering \summarization\ genomics

Architecture Overview

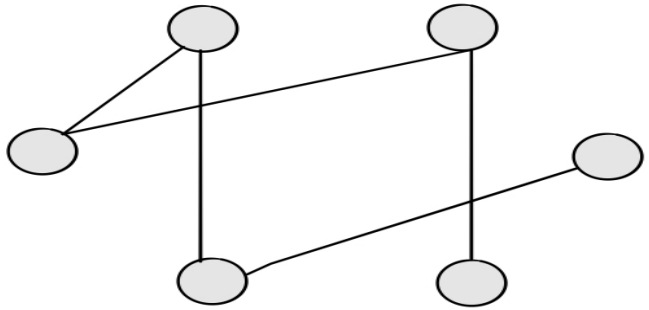


(a) Random attention (b) Window attention (c) Global Attention (d) BIGBIRD

Figure 1: Building blocks of the attention mechanism used in BIGBIRD. White color indicates absence of attention. (a) random attention with $r = 2$, (b) sliding window attention with $w = 3$ (c) global attention with $g = 2$. (d) the combined BIGBIRD model.

Random Attention

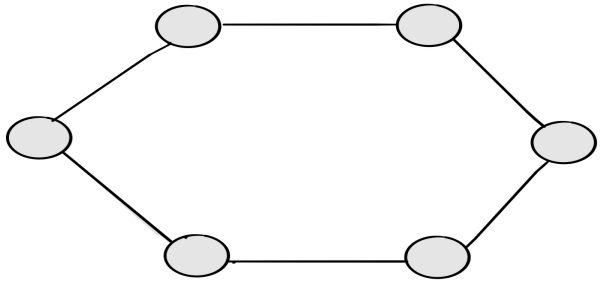
Each query attends over r random number of keys



- $O(n^2) \rightarrow O(r \cdot n) = O(n)$
- the shortest path between any two nodes is logarithmic in the number of nodes

Window Attention

A sliding window attention, so that during self attention of width w, query at location i attends from i – w/2 to i + w/2 keys.



- $O(n^2) \rightarrow O(w \cdot n) = O(n)$
- A great deal of information about a token can be derived from its neighboring tokens.
- Simple Erdos-Renyi random graphs do not have a high clustering coefficient, but small world graphs exhibit high clustering coefficient

Global Attention

- BigBird-ITC
  In internal transformer construction (itc), make some existing tokens “global”, which attend over the entire sequence
- BigBird-ETC
  In extended transformer construction (etc), include additional “global” tokens such as CLS

Theoretical Result

- Universal Approximators of sequence to sequence functions

Definition 1. *The star-graph S centered at 0 is the graph defined on $\{0, \ldots, n\}$. The neighborhood of all vertices i is $N(i) = \{0, i\}$ for $i \in \{1 \ldots n\}$ and $N(0) = \{1, \ldots n\}$.*

Theorem 1. *Given $1 < p < \infty$ and $\epsilon > 0$, for any $f \in \mathcal{F}_{CD}$, there exists a transformer with sparse-attention, $g \in \mathcal{T}_D^{H,m,q}$ such that $d_p(f, g) \leq \epsilon$ where D is any graph containing star graph S.*

- Turning Complete
- Limitations

We demonstrate a natural task which can be solved by the full attention mechanism in $O(1)$-layers. However, under standard complexity theoretic assumptions, this problem requires $\tilde{\Omega}(n)$-layers for any sparse attention layers with $\tilde{O}(n)$ edges (not just BIGBIRD). (Here $\tilde{O}$ hides poly-logarthmic factors.)

Experimental Results

QA

Model	HotpotQA			NaturalQ		TriviaQA	WikiHop
	Ans	Sup	Joint	LA	SA	Full	MCQ
RoBERTa	73.5	83.4	63.5	-	-	74.3	72.4
Longformer	74.3	84.4	64.4	-	-	75.2	75.0
BIGBIRD-ITC	**75.7**	86.8	67.7	70.8	53.3	**79.5**	**75.9**
BIGBIRD-ETC	75.5	**87.1**	**67.8**	**73.9**	**54.9**	78.7	**75.9**

Document Classification

Model	IMDb [65]	Yelp-5 [108]	Arxiv [36]	Patents [54]	Hyperpartisan [48]
# Examples	25000	650000	30043	1890093	645
# Classes	2	5	11	663	2
Excess fraction	0.14	0.04	1.00	0.90	0.53
SoTA	[89] 97.4	[3] 73.28	[70] 87.96	[70] 69.01	[41] 90.6
RoBERTa	95.0 ± 0.2	71.75	87.42	67.07	87.8 ± 0.8
BIGBIRD	95.2 ± 0.2	72.16	**92.31**	69.30	$\mathbf{92.2 \pm 1.7}$

Summarization

	Model	Arxiv			PubMed			BigPatent		
		R-1	R-2	R-L	R-1	R-2	R-L	R-1	R-2	R-L
Prior Art	SumBasic [69]	29.47	6.95	26.30	37.15	11.36	33.43	27.44	7.08	23.66
	LexRank [26]	33.85	10.73	28.99	39.19	13.89	34.59	35.57	10.47	29.03
	LSA [98]	29.91	7.42	25.67	33.89	9.93	29.70	-	-	-
	Attn-Seq2Seq [86]	29.30	6.00	25.56	31.55	8.52	27.38	28.74	7.87	24.66
	Pntr-Gen-Seq2Seq [78]	32.06	9.04	25.16	35.86	10.22	29.69	33.14	11.63	28.55
	Long-Doc-Seq2Seq [21]	35.80	11.05	31.80	38.93	15.37	35.21	-	-	-
	Sent-CLF [82]	34.01	8.71	30.41	45.01	19.91	41.16	36.20	10.99	31.83
	Sent-PTR [82]	42.32	15.63	38.06	43.30	17.92	39.47	34.21	10.78	30.07
	Extr-Abst-TLM [82]	41.62	14.69	38.03	42.13	16.27	39.21	38.65	12.31	34.09
	Dancer [32]	42.70	16.54	38.44	44.09	17.69	40.27	-	-	-
Base	Transformer	28.52	6.70	25.58	31.71	8.32	29.42	39.66	20.94	31.20
	+ RoBERTa [77]	31.98	8.13	29.53	35.77	13.85	33.32	41.11	22.10	32.58
	+ Pegasus [107]	34.81	10.16	30.14	39.98	15.15	35.89	43.55	20.43	31.80
	BIGBIRD-RoBERTa	41.22	16.43	36.96	43.70	19.32	39.99	55.69	37.27	45.56
Large	Pegasus (Reported) [107]	44.21	16.95	38.83	45.97	20.15	41.34	52.29	33.08	41.75
	Pegasus (Re-eval)	43.85	16.83	39.17	44.53	19.30	40.70	52.25	33.04	41.80
	BIGBIRD-Pegasus	**46.63**	**19.02**	**41.77**	**46.32**	**20.65**	**42.33**	**60.64**	**42.46**	**50.01**

Table 8: Summarization ROUGE score for long documents.

Genomics

- Promoter Region Prediction

Model	F1
CNNProm [91]	69.7
DeePromoter [72]	95.6
BIGBIRD	**99.9**

Table 10: Comparison.

- Chromatin-Profile Prediction

Model	TF	HM	DHS
gkm-SVM [31]	89.6	-	-
DeepSea [109]	95.8	85.6	**92.3**
BIGBIRD	**96.1**	**88.7**	92.1

Table 11: Chromatin-Profile Prediction

Conclusion

- BigBird satisfies all the known theoretical properties of full transformer
- the extended context modelled by BigBird greatly benefits variety of NLP tasks. (question answering and document summarization)
- introduce a novel application of attention based models where long contexts are beneficial: extracting contextual representations of genomics sequences like DNA.