

Big Bird: Transformers for Longer Sequences

Radek Bartyzal

GLAMI AI

1. 9. 2020

Motivation

What is old:

- Transformers are good
- self-attention is crucial for transformers
- self-attention = $O(n^2)$ memory + comp. requirements
- typical max sentence length = 512 tokens

What is new:

- is the $O(n^2)$ necessary?
- BigBird claims to replace self-attention with BigBird attention that is $O(n)$

Previous works

A lot of papers reducing the complexity of self-attention, notably:

- Synthesizer: synthesize the result of the full-attention
- Linformer: approximate full-attention by 2 low-rank matrices
- Longformer: Window + Global attention, 2 months earlier
- \implies BigBird = Longformer + Random attention

BigBird Attention

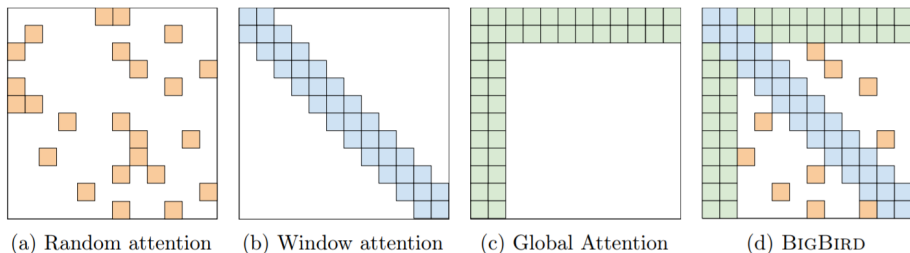


Figure: Building blocks of the attention mechanism used in BigBird. White color indicates absence of attention. (a) random attention with $r = 2$, (b) sliding window attention with $w = 3$ (c) global attention with $g = 2$. (d) the combined BigBird model.

- **Random:** each token attends to r random other tokens
- **Window:** all tokens attend to their surrounding tokens = similar to Convolution
- **Global:** certain tokens (e.g. CLS) receive from all and send to all

BigBird attention

- r, w, g is fixed = constant = hyperparameter
- therefore it is $O(n)$
- capable of simulating full self-attention
- the random mask is different for each sentence but fixed through the layers

Attention as graph

- Random attention = graph where each edge is independently chosen with a fixed probability.
- In such a random graph with just $\Theta(n)$ edges, the shortest path between any two nodes is logarithmic in the number of nodes.
- \implies quick mixing of information between nodes in following layers
- however in worst case it takes n layers to simulate 1 full self-attention layer
- \implies in worst case we would have to have $n \times$ more layers = $O(n^2)$ again
- \implies not truly $O(n)$ attention replacement

Global attention is necessary

Model	MLM	SQuAD	MNLI
BERT-base	64.2	88.5	83.4
Random (R)	60.1	83.0	80.2
Window (W)	58.3	76.4	73.1
R + W	62.7	85.1	80.5
Global + R + W	64.4	87.2	82.9

Table 1: Building block comparison @512

Figure: Random blocks and local window were insufficient in capturing all the context necessary to compete with the performance of BERT.

Cache optimization

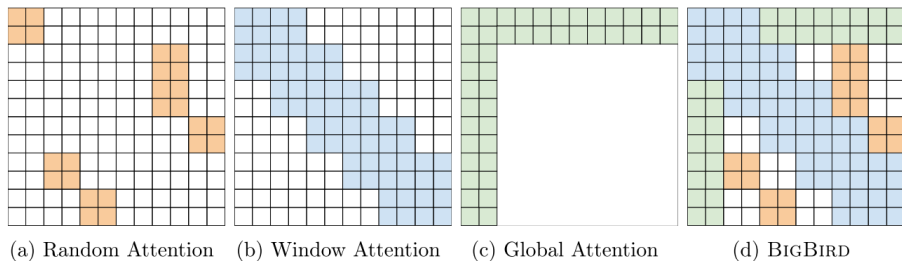


Figure: Building blocks of the block-attention mechanism used in BigBird with block size = 2. This implies the attention matrix is split into blocks of size 2×2 . All the previous BigBird parameters work on each block as a unit.

Question Answering results

Model	HotpotQA			NaturalQ		TriviaQA		WikiHop
	Ans	Sup	Joint	LA	SA	Full	Verified	MCQ
HGN [27]	82.2	88.5	74.2	-	-	-	-	-
GSAN	81.6	88.7	73.9	-	-	-	-	-
ReflectionNet [33]	-	-	-	77.1	64.1	-	-	-
RikiNet [62]	-	-	-	75.5	59.5	-	-	-
Fusion-in-Decoder [40]	-	-	-	-	-	84.5	90.3	-
SpanBERT [43]	-	-	-	-	-	79.1	86.6	-
MRC-GCN [88]	-	-	-	-	-	-	-	78.3
MultiHop [14]	-	-	-	-	-	-	-	76.5
Longformer [8]	81.2	85.8	73.2	-	-	77.3	85.3	81.9
BIGBIRD-ETC	81.2	89.1	73.6	77.7	57.8	80.9	90.8	82.3

Figure: Fine-tuning results on Test set for QA tasks. For Natural Questions Long Answer (LA), TriviaQA Verified, and WikiHop, **BigBird-ETC is the new state-of-the-art**.

Question Answering params = no random attention?

Parameter	HotpotQA	NaturalQ	TriviaQA	WikiHop
Global token location	ETC	ETC	ETC	ETC
# of global token, g	256	230	320	430
Window length, w	507	507	507	507
# of random token, r	0	0	0	0
Max. sequence length	4096	4096	4096	4096
# of heads	16	16	16	16
# of hidden layers	24	24	24	24
Hidden layer size	1024	1024	1024	1024
Batch size	32	64	32	64
Loss	cross-entropy	cross-entropy	cross-entropy	cross-entropy
Num epochs	{5, 9}	{3, 5}	{4, 6}	{5, 10}
Optimizer	Adam	Adam	Adam	LAMB
Learning rate	3×10^{-5}	$\{5, 10\} \times 10^{-5}$	$\{3, 5\} \times 10^{-5}$	$\{2, 5\} \times 10^{-5}$
Compute resources	4×4 TPUv3	4×8 TPUv3	4×4 TPUv3	4×8 TPUv3

Figure: Hyperparameters of large BigBird model for Question Answering submitted for test.

Conclusion

- a lot of engineering to optimize Window + Global + Random attention
- allows longer sequence length and larger batch size
- \implies which brings improvements on almost all tasks

Sources

1. Zaheer, Manzil, et al. "Big bird: Transformers for longer sequences." arXiv preprint arXiv:2007.14062 (2020).
<https://arxiv.org/abs/2007.14062v1>