Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh et al.

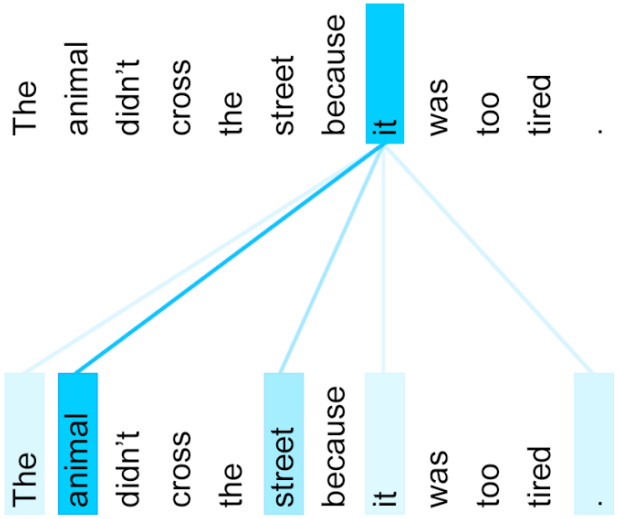
Google

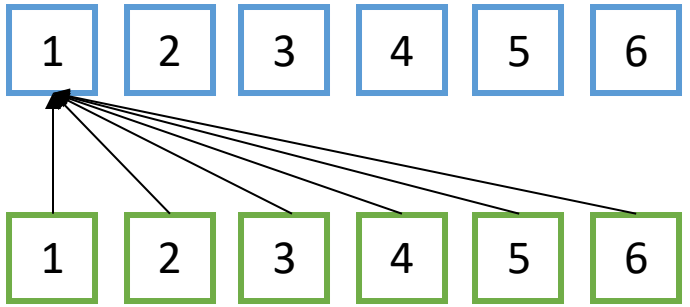
Notable Works

- Longer length work-around using sliding window: **SpanBERT**, **ORQA**, **REALM**, **RAG**, etc.
- Reducing full-attention: **Child et al**. reduces complexity to O(N$\sqrt{N}$), **Kitav et al.** O(N log(N))
- **Longformer** and **Extended Transformers Construction**
- Understanding Self-Attention
  - Expressivity **Yun et al.**
  - Turing Complete **Perez et al.**

Problems With BERT (and variants)

- Dependence on self-attention



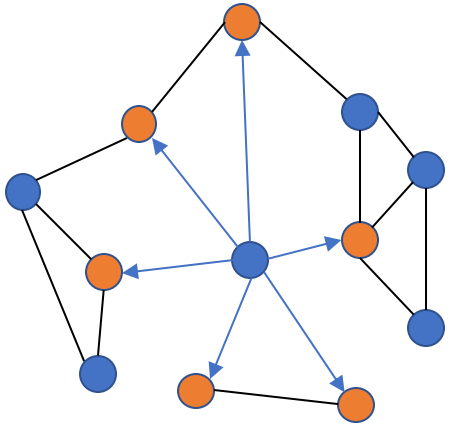


- Quadratic dependency (mainly memory) on the sequence length.

Claims

- O(n) inner-products. Quadratic Dependency to Linear
- New SOTA for question answering and summarization (longer sequences) The extended context (4096 tokens) greatly helps the tasks.
- Genomics sequences applications (novel). Longer masked LM pretraining helps DNA-based tasks.

Generalized Attention Mechanism



$$\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n) \in \mathbb{R}^{n \times d}$$

D: Di-graph with vertex set is [n] = {1, …, n}

$N(i)$: Out-neighbors set of node i

$$\text{ATTN}_D(\boldsymbol{X})_i = \boldsymbol{x}_i + \sum_{h=1}^{H} \sigma\left(Q_h(\boldsymbol{x}_i) K_h(\boldsymbol{X}_{N(i)})^T\right) \cdot V_h(\boldsymbol{X}_{N(i)})$$

Generalized Attention Mechanism

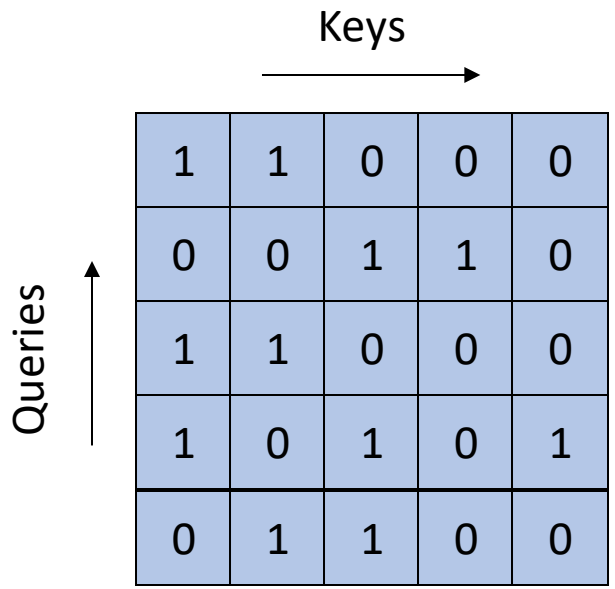
$$\textsc{Attn}_D(\boldsymbol{X})_i = \boldsymbol{x}_i + \sum_{h=1}^{H} \sigma\left(Q_h(\boldsymbol{x}_i) K_h(\boldsymbol{X}_{N(i)})^T\right) \cdot V_h(\boldsymbol{X}_{N(i)})$$

$$Q_h, K_h : \mathbb{R}^d \to \mathbb{R}^m \qquad V_h : \mathbb{R}^d \to \mathbb{R}^d$$

$X_{N(i)}$: Stacking $\quad \{\boldsymbol{x}_j : j \in N(i)\}$

Generalized Attention Mechanism

- Adjacency Matrix A(i, j) = 1
- Query-i and Key-j



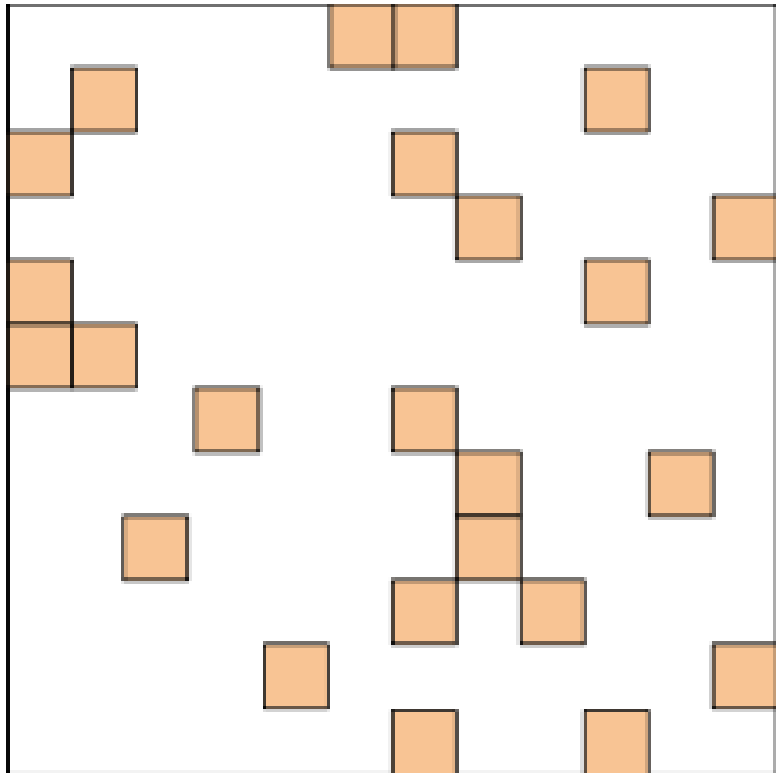
The Architecture

Sparse Attention

- Graph sparsification problem: Random graphs can approximate complete graphs (expanders)
- Erdős-Rényi model: Choose edges independently with some prob. With Θ(n) edges, shortest path is logarithmic. Approximates complete graph.

BigBird has **sparse attention**: A(i,.) = 1 for r keys.

Sparse/Random Attention



(a) Random attention

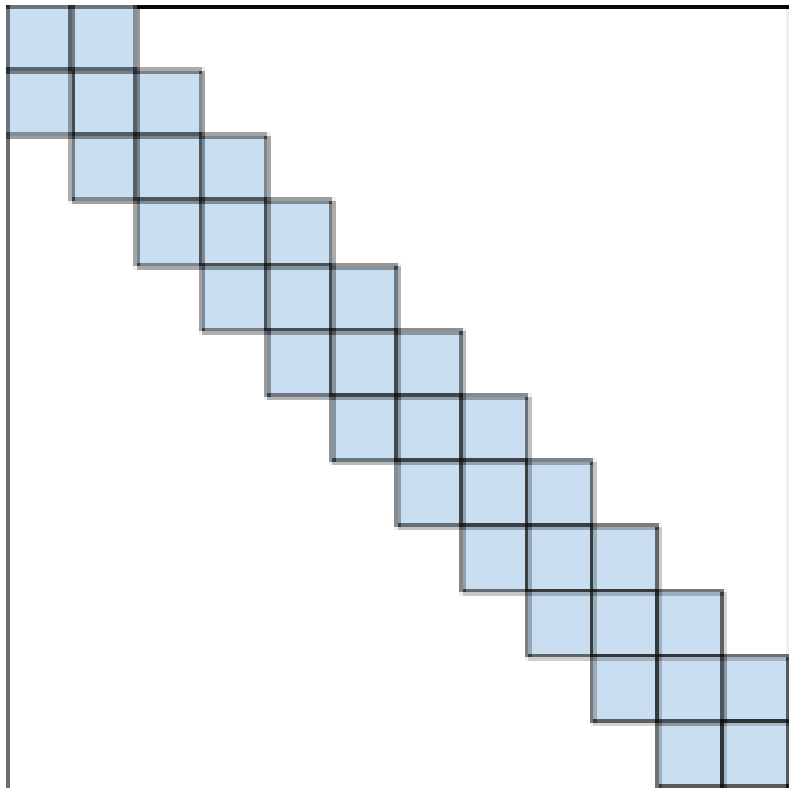
The Architecture

Locality Of Reference

- Graph theory: Clustering coefficient high when there are cliques or near-cliques. Erdős-Rényi model lacks this.
- World graphs exhibit this property.

BigBird has **world attention**: $A(i, i - w/2 : i + w/2) = 1$

Window Attention



(b) Window attention

The Architecture

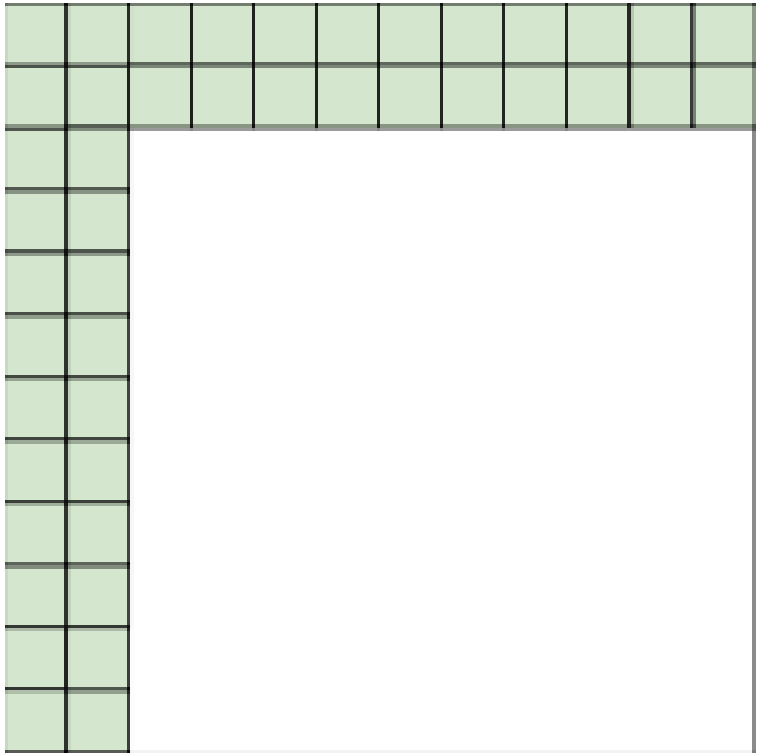
Global Attention

- Just Random attn and Window attn are not enough.
- Global tokens: Attend to all other tokens, and all other tokens attend them.
- Two different models:
  - BIGBIRD-ITC (Internal Transformer Construction)
  - BIGBIRD-ETC (External Transformer Construction)

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

Global Attention

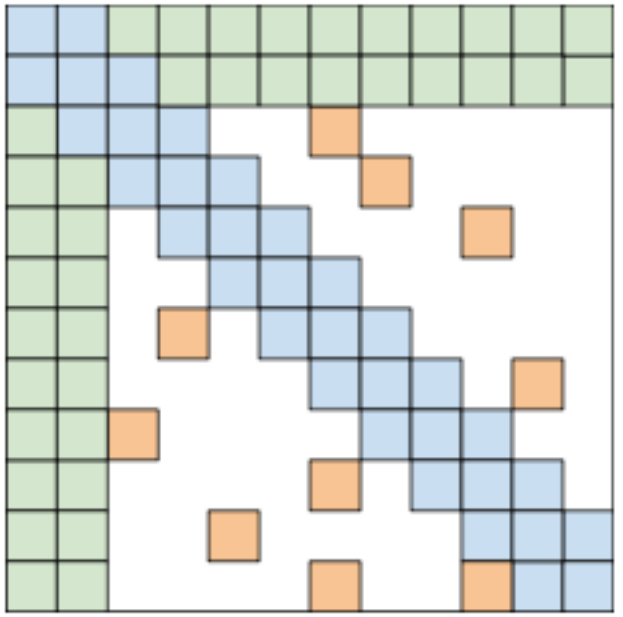


(c) Global Attention

BigBird

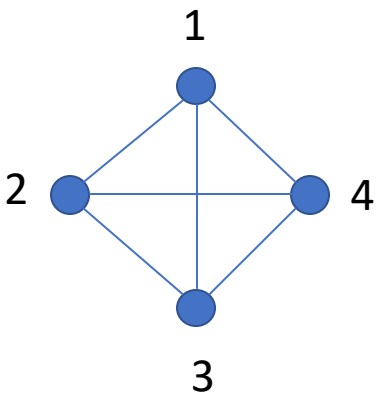
Final Architecture

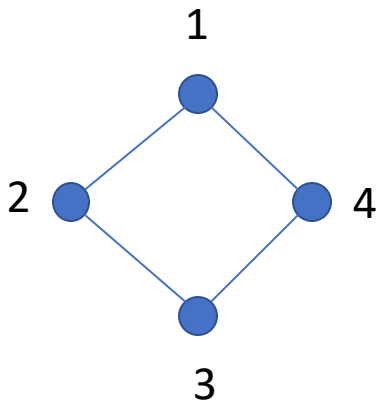
1. Queries attend to random keys
2. Locality
3. Global Tokens



(d) BIGBIRD

Idea





Theoretic Arguments

- Universal Approximators

- Turing Completeness

Following the arguments in Yun et al (Expressivity) and Perez et al (Turing Complete)

Limitations

Consider the problem:

Given n unit vectors $\{u_1, \ldots, u_n\}$, find $f(u_1, \ldots, u_n) \to (u_{1^*}, \ldots, u_{n^*})$

$j^* = \arg\max_k \|u_k - u_j\|_2^2$

- A single full-attention layer solves this in O(1)
- However, sparse attention with O(n) edges requires Ω(n)-layers

Experiments

Pretraining (MLM)

- Predict random subset of masked-out tokens.
- Warm-starting using RoBERTa checkpoint.
- Performance of masked-out token prediction (using bits per character)

$$bpc(string) = \frac{1}{T}\sum_{t=1}^{T} H(P_t, \hat{P}_t) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{c=1}^{n} P_t(c)\log_2 \hat{P}_t(c),$$

Model	Base	Large
RoBERTa (sqln: 512)	1.846	1.496
Longformer (sqln: 4096)	1.705	1.358
BIGBIRD-ITC (sqln: 4096)	1.678	1.456
BIGBIRD-ETC (sqln: 4096)	**1.611**	**1.274**

Experiments: Question Answering

Baseline

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**

Table 4: QA Dev results using Base size models. We report accuracy for WikiHop and F1 for HotpotQA, Natural Questions, and TriviaQA.

Leaderboard

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**

Experiments: Document Classification

- Excess fraction: fraction of dataset that exceeds 512 tokens.
- New SOTA on Arxiv dataset.
- Does not outperform if excess fraction is less.

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	$\mathbf{92.2 \pm 1.7}$

Table 6: Classification results. We report the F1 micro-averaged score for all datasets. Experiments on smaller IMDb and Hyperpartisan datasets are repeated 5 times and the average performance is presented along with standard deviation.

Experiments: Encoder-Decoder

- The encoder uses BigBird sparse attention. Decoder uses full attention.
- MLP pretraining on base size model
- Pegasus pre-training on large-sized BigBird model.

	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.

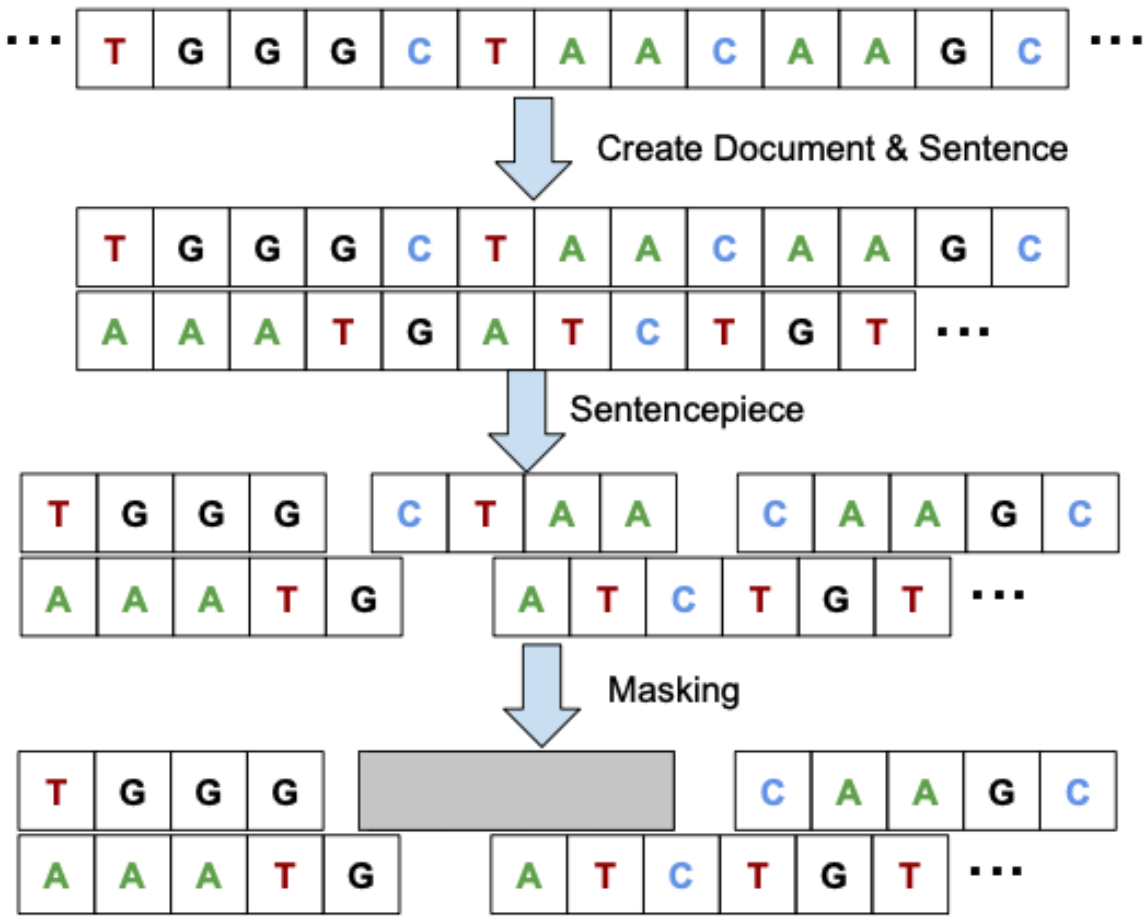
	Instances			Input Length		Output Length	
Dataset	Training	Dev	Test	Median	90%-ile	Median	90%-ile
Arxiv [21]	203037	6436	6440	6151	14405	171	352
PubMed [21]	119924	6633	6658	2715	6101	212	318
BigPatent [79]	1207222	67068	67072	3082	7693	123	197

Table 7: Statistics of datasets used for summarization.

Results: Genomics

MLP Pretraining

- Split GRCh37 DNA seq split at random positions. (Documents). 50-100 sentences (500-1000 bps each)
- This repeated 10 times for each chromosome.
- Final tokens contain 8bps on average.
- 10% tokens are masked and predicted.



Results: Genomics

- Promoter Region Prediction
  - Task: Classify DNA sequence as promoter or non-promoter.
  - Eukaryotic Promoter Database.

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

- Chromatin-Profile Prediction
  - Task: Predict non-chromatin profile in non-coding region of DNA.
  - 919 binary classifiers for 919 chromatin profiles

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

Hyperparameters

Parameter	BIGBIRD-ITC	BIGBIRD-ETC
Block length, b	64	84
# of global token, g	$2 \times b$	256
Window length, w	$3 \times b$	$3 \times b$
# of random token, r	$3 \times b$	0
Max. sequence length	4096	4096
# of heads	12	12
# of hidden layers	12	12
Hidden layer size	768	768
Batch size	256	256
Loss	MLM	MLM
Activation layer	gelu	gelu
Dropout prob	0.1	0.1
Attention dropout prob	0.1	0.1
Optimizer	Adam	Adam
Learning rate	10^{-4}	10^{-4}
Compute resources	8×8 TPUv3	8×8 TPUv3

Table 12: Hyperparameters for the two BIGBIRD base models for MLM.

Conclusion

- Sparse attention mechanism that is linear in # tokens.
- Theoretical results: Universal Approximators and Turing Complete.
- SOTA on several NLP tasks.
- Attention based contextual LM for DNA. SOTA on downstream tasks.