Big Bird : Transformers for Longer Sequences

Zaheer, Manzil, et al.,

34th Conference on Neural Information Processing Systems (NeurIPS. 2020)

Seungone Kim

Department of Computer Science

Yonsei University

louisdebroglie@yonsei.ac.kr

2021.07.28

연세대학교 YONSEI UNIVERSITY | SOFT COMPUTING LABORATORY

Outline

- 문제 정의
- 관련 연구
- 제안하는 방법
- 실험 결과
- 결론

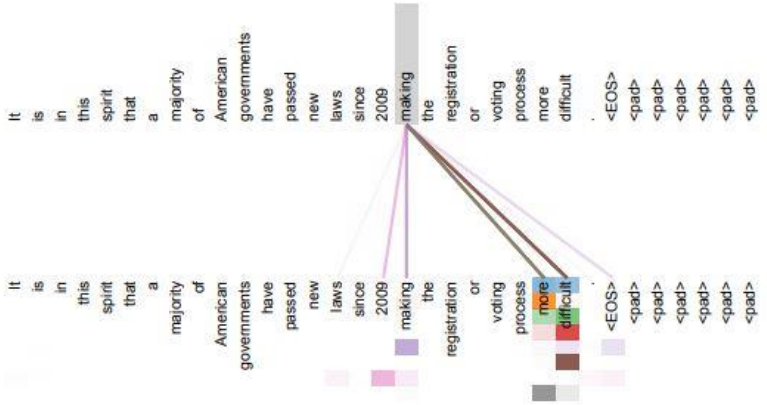
문제 정의

- Transformers-based models have quadratic dependency on the sequence length
  - This is due to full attention mechanism
  - With self-attention mechanism, models achieved versatility and robustness
  - Self attention mechanism eliminated sequential dependency in RNNs (sequential nature of RNNs)
  - Parallelism enables Transformers to leverage the full power of modern GPUs/TPUs

- What aspects of self-attention model are essential for its performance?
  - Can we achieve preserve the expressivity and flexibility by using fewer inner-products?
  - Reduce memory and computation requirements

Table 1: Maximum path lengths, per-layer complexity and minimum number of sequential operations for different layer types. n is the sequence length, d is the representation dimension, k is the kernel size of convolutions and r the size of the neighborhood in restricted self-attention.

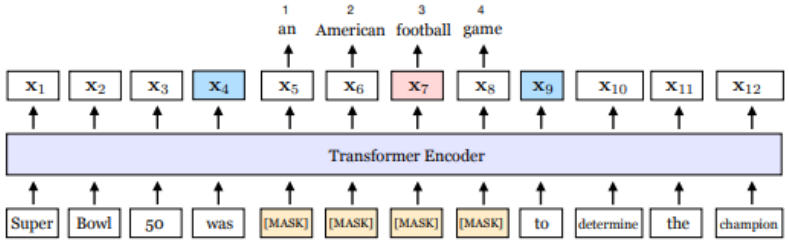
Layer Type	Complexity per Layer	Sequential Operations	Maximum Path Length
Self-Attention	$O(n^2 \cdot d)$	$O(1)$	$O(1)$
Recurrent	$O(n \cdot d^2)$	$O(n)$	$O(n)$
Convolutional	$O(k \cdot n \cdot d^2)$	$O(1)$	$O(log_k(n))$
Self-Attention (restricted)	$O(r \cdot n \cdot d)$	$O(1)$	$O(n/r)$



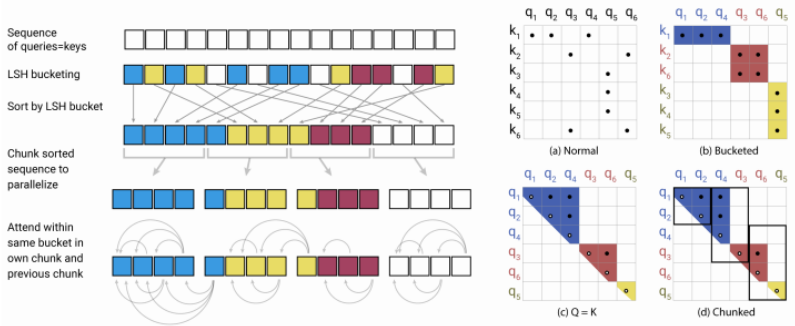
관련 연구

- Research alleviating quadratic dependency of Transformers
  - We can categorize into two directions

  - 1) Embrace the length limitation and develop method around it
    - Using some other mechanism to select smaller subset of relevant contexts to feed to model and iterate
    - e.g. Call Transformer Block multiple times with different contexts each time
    - SpanBERT, ORQA, REALM, RAG

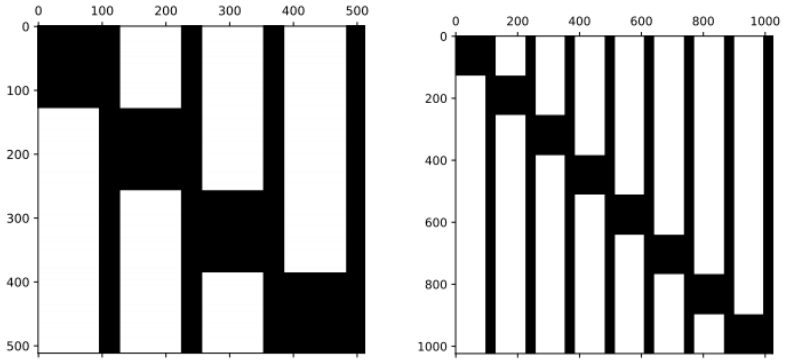


  - 2) Come up with methods that do not require Full Attention
    - Child et al. "Generating long sequences with sparse transformers." proposes O(n$\sqrt{n}$) complexity method
    - Kitaev et al. "Reformer: The Efficient Transformer." proposes O(n$\log n$) complexity method with LTC(hashing)

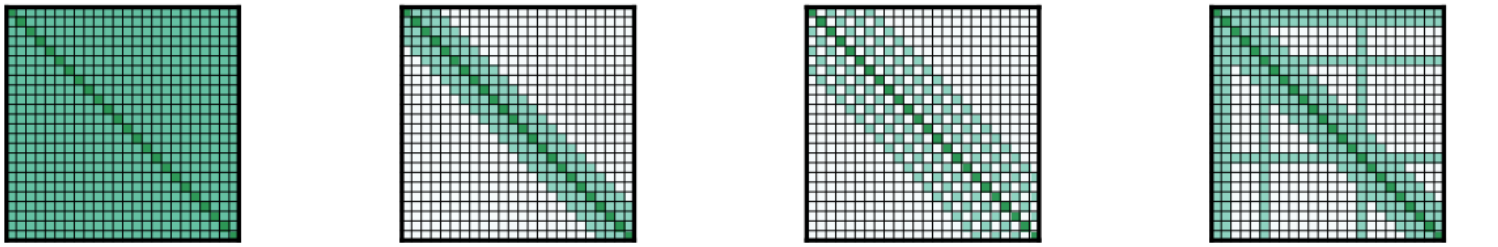


관련 연구

- Quadratic Complexity of Self-Attention limits its application on long text
  - *Qiu et al. "Blockwise Self-Attention for Long Document Understanding" (EMNLP 2020)* proposed using block sparsity to reduce complexity



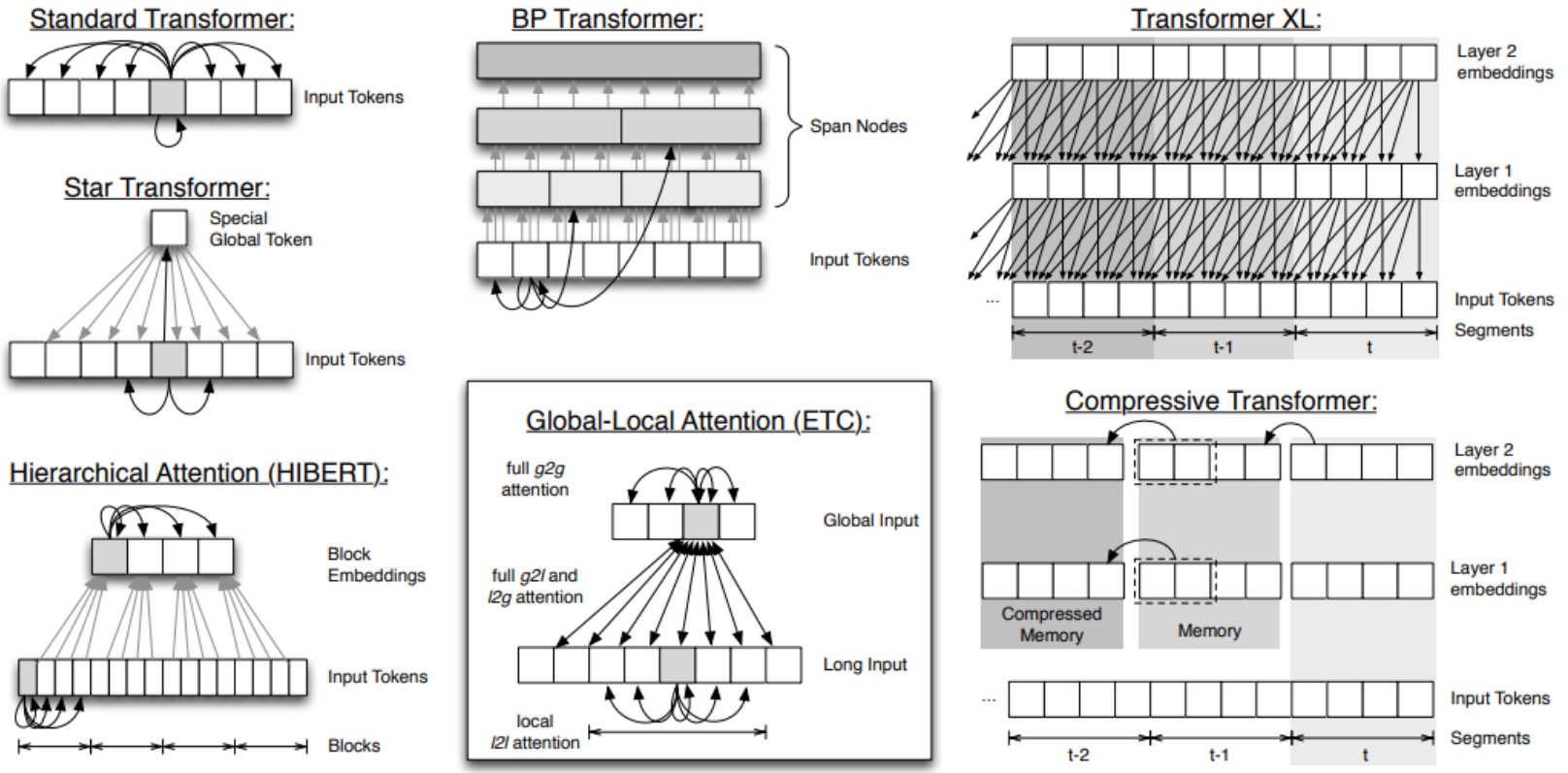
  - *Beltagy et al. "Longformer: The long-document transformer."* proposed using a localized sliding window based mask with few global mask to reduce computation



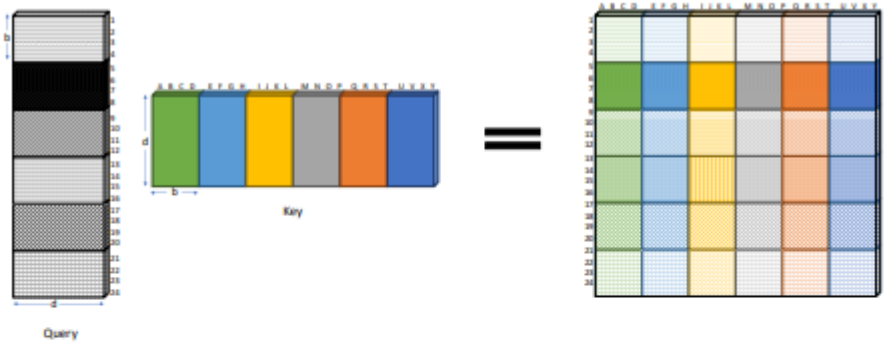
(a) Full n^2 attention (b) Sliding window attention (c) Dilated sliding window (d) Global+sliding window

관련 연구

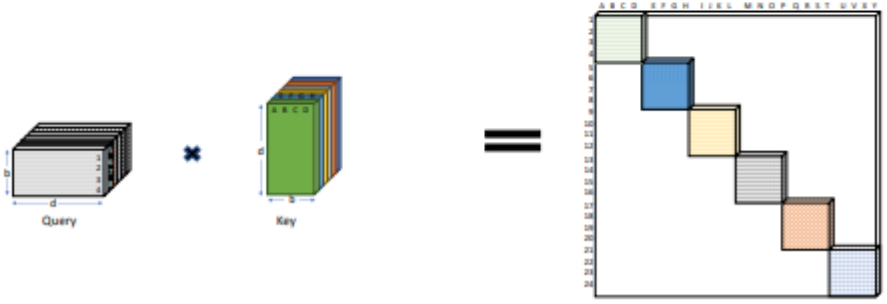
- Quadratic Complexity of Self-Attention limits its application on long text
  - *Ainslie et al., "ETC : encoding long and structured data in transformers" (EMNLP 2020)* proposed a method to encode structure in text for Transformers
  - The idea of global tokens was used extensively to achieve their goals



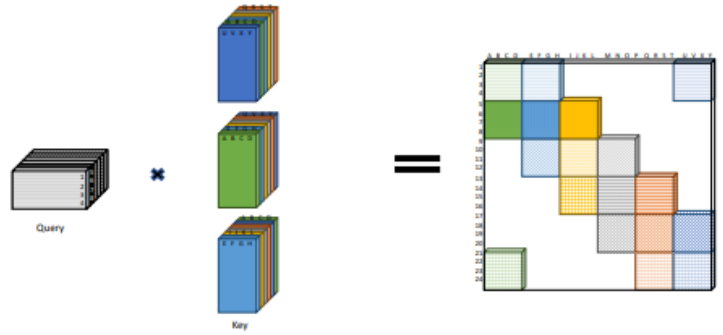
Attention Visualization



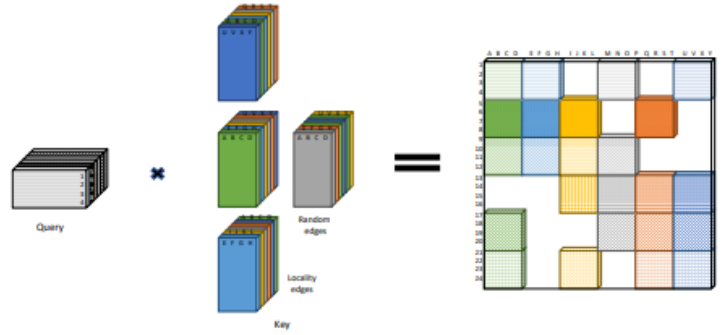
(a) Full all pair attention can be obtained by direct matrix multiplication between the query and key matrix. Groupings just shown for guidance.



(b) Block diagonal attention can be computed by "blockifying" the query and key matrix



(c) Window local attention obtained by "blockifying" the query/key matrix, copying key matrix, and rolling the resulting key tensor (Obtaining rolled key-block tensor is illustrated in detail in Fig. 5). This ensures that every query attends to at least one block and at most two blocks of keys of size b on each side.



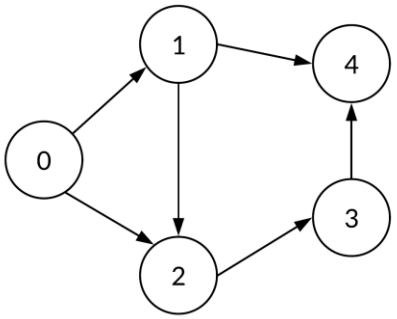
(d) Window + Random attention obtained by following the procedure above along with gathering some random key blocks.

제안하는 방법

- Generalised Attention Mechanism
  - Described by a directed graph D whose vertex set is $[n] = \{1,2,\dots,n\}$
  - The set of arcs(directed) represent the set of inner products attention mechanism will consider

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

  - Q_h, K_h: $\mathbb{R}^d \to \mathbb{R}^d$ are query and key functions
  - V_h: $\mathbb{R}^d \to \mathbb{R}^d$ is a value function
  - σ is a scoring function (softmax)
  - H denotes the number of heads
  - $X_{N(i)}$ corresponds to the matrix formed by stacking $\{x_j : j \in N(i)\}$ and not all inputs
  - If D is a complete digraph, we recover the full quadratic attention mechanism
  - Operate on adjacency matrix A of graph D even though graph maybe sparse

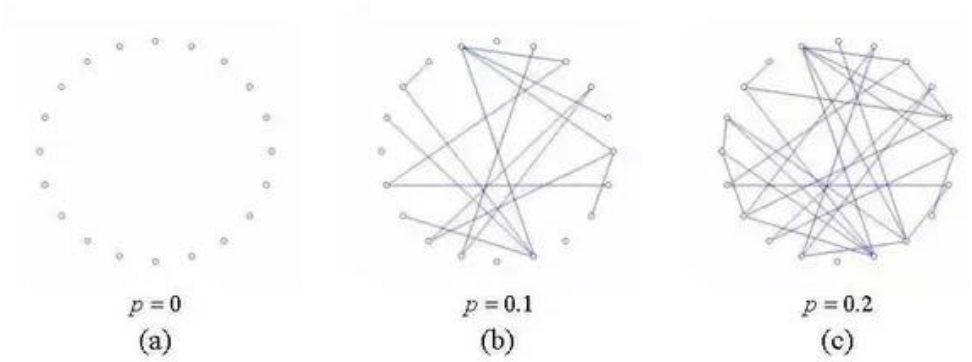


Adjacency Matrix

	0	1	2	3	4
0	0	1	1	0	0
1	0	0	1	0	1
2	0	0	0	1	0
3	0	0	0	0	1
4	0	0	0	0	0

제안하는 방법

- Graph Sparsification Problem
  - It is well known that random graphs are expanders and can approximate complete graphs in a number of different contexts including their spectral properties
    - *Spielman, Daniel A., and Shang-Hua Teng. "Spectral sparsification of graphs." SIAM Journal on Computing 40.4 (2011): 981-1025.*
    - *Hoory, Shlomo, Nathan Linial, and Avi Wigderson. "Expander graphs and their applications." Bulletin of the American Mathematical Society 43.4 (2006): 439-561.*
  - Sparse Random Graph for Attention Mechanism have two desiderata
    - Small Average Path length between nodes
    - Notion of Locality
  - Consider the simplest Random Graph Construction
    - Erdős–Rényi model
    - A graph on n vertices such that the presence of each edge is p



제안하는 방법

- Graph Sparsification Problem
  - Erdős–Rényi model
  - In a random graph with just n edges, the shortest path between two nodes is logarithmic in number of nodes
    - *Chung, Fan, and Linyuan Lu. "The average distances in random graphs with given expected degrees." Proceedings of the National Academy of Sciences 99.25 (2002): 15879-15882.*
    - *Katzav, Eytan, Ofer Biham, and Alexander K. Hartmann. "Distribution of shortest path lengths in subcritical Erdős-Rényi networks." Physical Review E 98.1 (2018): 012301.*
  - The second eigen value of a adjacency matrix is quite far from first eigen value
    - *Alt, Johannes, Raphaël Ducatez, and Antti Knowles. "Extremal eigenvalues of critical Erdős–Rényi graphs." The Annals of Probability 49.3 (2021): 1347-1401.*
    - *Benaych-Georges, Florent, Charles Bordenave, and Antti Knowles. "Spectral radii of sparse random matrices." Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. Vol. 56. No. 3. Institut Henri Poincaré, 2020.*
    - *Benaych-Georges, Florent, Charles Bordenave, and Antti Knowles. "Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs." The Annals of Probability 47.3 (2019): 1653-1676.*
    - This property leads to a rapid mixing time for random walks in the graph, meaning that information can flow fast between any pair of nodes

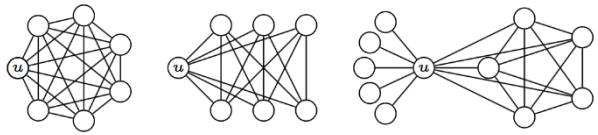
제안하는 방법

- Graph Sparsification Problem

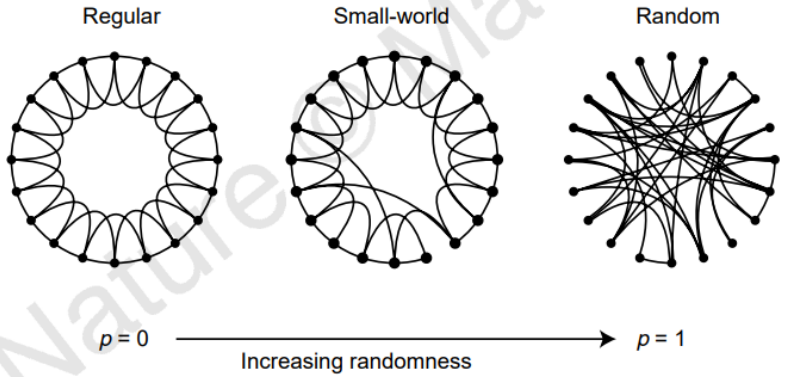
  - A great deal of information about a token can be derived from its neighboring tokens
    - *Clark, Kevin, et al. "What Does BERT Look at? An Analysis of BERT's Attention." Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 2019.*

  - In graph theory, Clustering Coefficient is a measure of locality connectivity
    - Value is high when graph contains many sub-graphs that are almost fully interconnected

  - Erdős–Rényi model does not have high clustering coefficient
    - *Watts, Duncan J., and Steven H. Strogatz. "Collective dynamics of 'small-world'networks." nature 393.6684 (1998): 440-442.*
    - Good balance between average shortest path and notion of locality
    - Goal is to construct a graph with n nodes each connected to w neighbors, w/2 on each side
    - Begin with a Sliding Window on the nodes

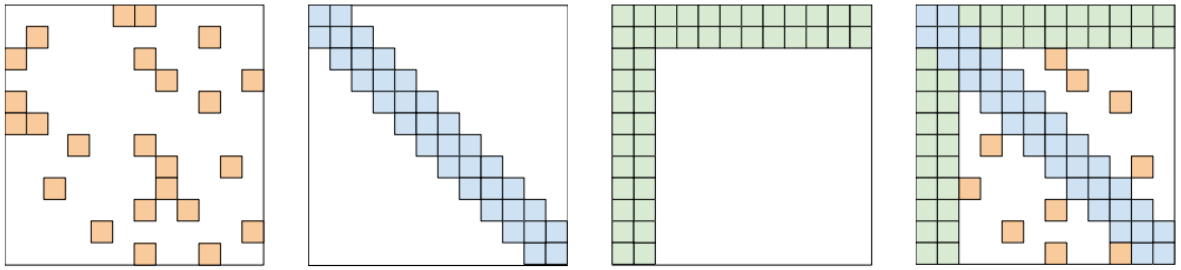


$C_2(u)$	1	$\frac{d}{2(d-1)} \approx \frac{1}{2}$	$\frac{d-2}{4d-4} \approx \frac{1}{4}$
$C_3(u)$	1	0	$\frac{d-4}{2d-4} \approx \frac{1}{2}$
$C_4(u)$	1	0	$\frac{d-6}{2d-6} \approx \frac{1}{2}$



제안하는 방법

- Random Blocks + Local Window + Global Tokens
  - Random Blocks and Local Window were insufficient in capturing all the context necessary to compete with the performance of BERT
  - Final piece of BIGBIRD is to utilize global tokens
  - Global Tokens attend to all tokens in the sequence
    - BIGBIRD-ITC (Internal Transformer Construction) : make some existing tokens global
    - BIGBIRD-ETC (Extended Transformer Construction) : include additional global tokens such as CLS
  - Final Attention mechanism for BIGBIRD
    - Queries attend to r random keys
    - Queries attend to w/2 tokens to left and w/2 tokens to right
    - Contain g global tokens



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

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

Table 1: Building block comparison @512

Attention Visualization

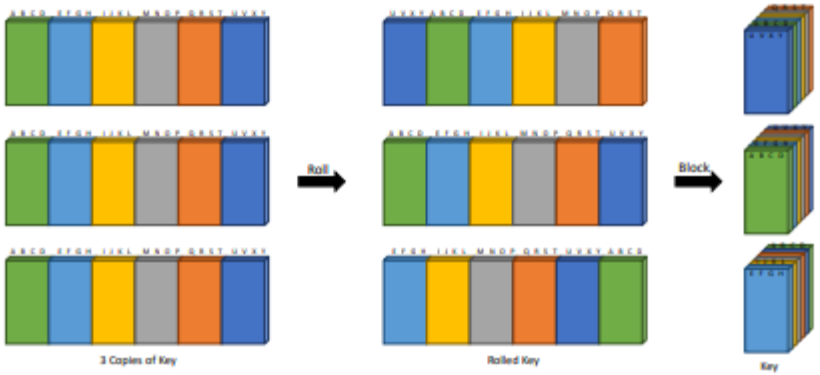


Figure 5: Construction of rolled key-block tensor. Make w copies of the key matrix. Index the copies as $-(w-1)/2 \leq j \leq (w-1)/2$. Roll j^{th} copy by j blocks. Positive roll means circular shift entries left and likewise for negative roll corresponds to right shift. Finally, reshape by grouping the blocks along a new axis to obtain the key-blocked tensor. For illustration purpose $w = 3$ is chosen.

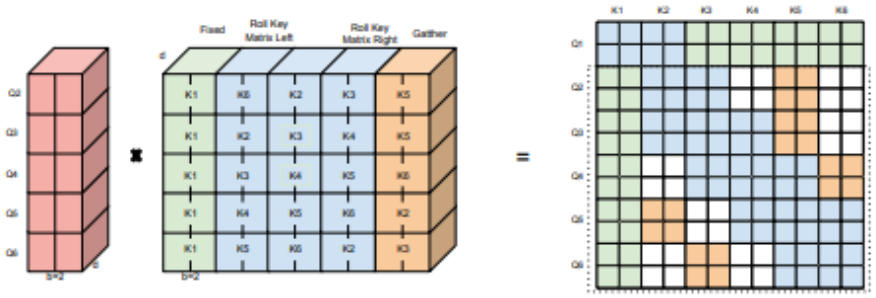


Figure 6: Overview of BIGBIRD attention computation. Structured block sparsity helps in compactly packing our operations of sparse attention, thereby making our method efficient on GPU/TPU. On the left, we depict the transformed dense query and key tensors. The query tensor is obtained by simply blocking and reshaping while the final key tensor by concatenating three transformations: The first green columns, corresponding to global attention, is fixed. The middle blue columns correspond to window local attention and can be obtained by appropriately rolling as illustrated in Fig. 5. For the final orange columns, corresponding to random attentions, we need to use computationally inefficient gather operation. Dense multiplication between the query and key tensors efficiently calculates the sparse attention pattern (except the first row-block, which is computed by direct multiplication), using the ideas illustrated in Fig. 4. The resultant matrix on the right is same as that shown in Fig. 3d.

제안하는 방법

- Are Sparse Attention powerful as Full Attention?
  - Sparse Attention Mechanisms used in stand-alone Encoder(e.g. BERT), they are Universal Approximators of Sequence to Sequence functions
    - *Yun, Chulhee, et al. "Are Transformers universal approximators of sequence-to-sequence functions?." International Conference on Learning Representations. 2019.*
    - *Yun, Chulhee, et al. "O (n) Connections are Expressive Enough: Universal Approximability of Sparse Transformers." NeurIPS. 2020.*
  - Sparse Encoder-Decoder Transformers are Turing Complete
    - *Pérez, Jorge, Javier Marinković, and Pablo Barceló. "On the Turing Completeness of Modern Neural Network Architectures." International Conference on Learning Representations. 2018.*
  - Moving to sparse-attention mechanism incurs a cost
    - No Free Lunch
    - Sparse Mechanism will require polynomially more layers
  - Key difference between BIGBIRD's attention mechanism to that of *Vaswani et al., Yun et al* is that the authors added a special token at the beginning of each sequence and assign it to a special vector
    - Therefore graph D will have vertex set $\{0\} \cup [n] = \{0, 1, 2, \ldots, n\}$
    - Extra node and its respective vector will be dropped at final output layer

제안하는 방법

- Universal Approximators
  - Sparse Attention Mechanism defined by any graph containing S is a universal approximator

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.*

  - To prove the above theorem, authors followed proof structure from *Yun et al., "Are Transformers universal approximators of sequence-to-sequence functions?." (ICLR 2019)*

  - *Approximate* F_{CD} by piece-wise constant function
  - Approximate piece-wise constant functions by modified transformers
  - Approximate modified transformers by original Transformers

  - Creating a contextual mapping with sparse attention is quite a challenge because each query only attends to a few keys, it is not all clear sufficient information can be corralled to make a contextual embedding of the entire matrix
  - Authors developed sparse shift operator and used global tokens to handle this problem

제안하는 방법

- Universal Approximators

Lemma 1 (Lemma 8 [104]). *For any given $f \in \mathcal{F}_{CD}$ and $1 \leq p \leq \infty$, there exists a $\delta > 0$ such that there exists a piece-wise constant function $\bar{f}$ with $d_p(f, \bar{f}) \leq \frac{\epsilon}{3}$. Concretely, $\bar{f}$ is defined as*

$$\bar{f}(X) = \sum_{P \in G_\delta} f(P) \cdot \mathbf{1}\left[\|\mathrm{ReLU}(X - P)\|_\infty \leq \delta\right]$$

Lemma 2. *Given a function $\psi : \mathbb{R}^{(n+1)\times(d+1)} \times \mathbb{R}^2 \to \mathbb{R}^{(n+1)\times 1}$ and a vector $u \in \mathbb{R}^{d+1}$ and a sparse attention mechanism based on the directed graph D, we can implement a selective shift operator that receives as input a matrix $X \in \mathbb{R}^{(n+1)\times(d+1)}$ and outputs $X + \rho \cdot \psi_u(X, b_1, b_2)$ where*

$$\psi_u(Z; b_1, b_2)_i = \begin{cases} (\max_{j \in N(i)} u^T Z_j - \min_{j \in N(i)} u^T Z_j) e_1 & \text{if } b_1 \leq u^T Z_j \leq b_2 \\ 0 & \text{else.} \end{cases}$$

Note that $e_1 \in R^{d+1}$ denotes $(1, 0, \ldots, 0)$.

Lemma 3. *There exists a function $g_c : \mathbb{R}^{(n+1)\times(d+1)} \to \mathbb{R}^{(n+1)}$ and a unique vector u, such that for all $P \in \mathbf{G}_\delta^E$ $g_c(P) := \langle u, g(P) \rangle$ satisfies the property that g_c is a contextual mapping of P. Furthermore, $g_c \in \mathcal{T}_D^{2,1,1}$ using a composition of sparse attention layers as long as D contains the star graph.*

Lemma 4 (Lemma 7 [104]). *Let g_c be the function in Lemma 3, we can construct a function $g_v : \mathbb{R}^{(n+1)\times(d+1)} \to \mathbb{R}^{(n+1)\times d}$ composed of $O(n\delta^{-nd})$ feed-forward layers (with hidden dimension $q = 1$) with activations in Φ such that g_v is defined as $g_v(Z) = [g_v^{tkn}(Z_1), \ldots, g_v^{tkn}(Z_n)]$, where for all $j \in \{1, \ldots, n\}$,*

$$g_v^{tkn}(g_c(L)_j) = f(L)_j$$

Lemma 5 (Lemma 9 [104]). *For each $g \in \bar{\mathcal{T}}^{2,1,1}$ and $1 \leq p \leq \infty$, $\exists g \in \mathcal{T}^{2,1,4}$ such that $d_p(g, \bar{g}) \leq \epsilon/3$*

제안하는 방법

- Turing Completeness
  - What the additional power of both a decoder along with an encoder is?
  - *Pérez, Jorge, Javier Marinković, and Pablo Barceló. "On the Turing Completeness of Modern Neural Network Architectures."(LREC 2018)* showed full transformer based on a quadratic attention mechanism is Turing Complete
  - Use a Sparse Encoder and Decoder to simulate any Turing Machine
  - Turing machine can read and write symbols on an infinite tape according to some rules, in other words, it can perform any computation / algorithm on it no matter how complicated it is
    - https://www.cl.cam.ac.uk/projects/raspberrypi/tutorials/turing-machine/one.html (University of Cambridge)

Turning Machine We will use the same setup of Turning Machine that was used by Pérez et al. [72] (see section B.4). Given a Turing Machine $M = (Q, \Sigma, \delta, q_{init}, F)$, we use the following notation

$q^{(j)}$: state of Turing machine M at time j.

$s^{(j)}$: symbol under the head of M at time j.

$v^{(j)}$: symbol written by M at time j.

$m^{(j)}$: head direction in the transition of M at time j.

Lemma 6 (Lemma B.2 [72]). *There exists a two-layer feed-forward network $O_1 : \mathbb{Q}^d \to \mathbb{Q}^d$ such that with input vector $\boldsymbol{a}_i^1$ (Eq. (7)) produces as output*

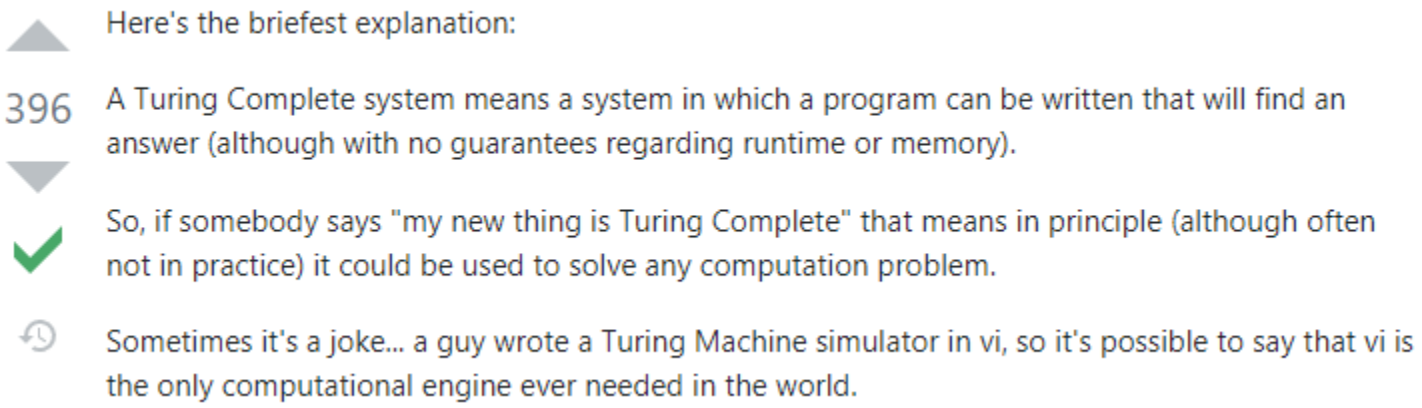
$$O_1(\boldsymbol{a}_i^1) = [\ 0, \ldots, 0, \ [\![q^{g(i)+1}]\!], [\![v^{g(i)}]\!], m^{g(i)}, 0, 0, 0, 0 \ 0, \ldots, 0, \ 0, \ldots, 0 \ ]$$

Lemma 7 (Lemma B.5 [72]). *There exists a function $F : \mathbb{Q}^d \to \mathbb{Q}^d$ defined by a feed-forward network such that*

$$F(\boldsymbol{z}_r^4) = [\ [\![q^{g(r+1)}]\!], [\![s^{g(r+1))}]\!], c^{g(r+1)}, \ 0, \ldots, 0, \ \mathbf{0}_s, 0, [\![w^{(r+1)}]\!], \ 0, 0, 0, 0, 0, u_1^{(r+1)}, u_2^{(r+1)}, u_3^{(r+1)}, u_4^{(r+1)}]$$
$$= \boldsymbol{y}_{r+1}$$

제안하는 방법

- What is Turing Complete / Turing Complete



396

Here's the briefest explanation:

A Turing Complete system means a system in which a program can be written that will find an answer (although with no guarantees regarding runtime or memory).

So, if somebody says "my new thing is Turing Complete" that means in principle (although often not in practice) it could be used to solve any computation problem.

Sometimes it's a joke... a guy wrote a Turing Machine simulator in vi, so it's possible to say that vi is the only computational engine ever needed in the world.

제안하는 방법

- Limitations
  - Problem requires n more layers for any sparse attention layers with O(n) edges

Task 1. Given n unit vectors $\{u_1, \ldots, u_n\}$, find $f(u_1, \ldots, u_n) \to (u_{1^*}, \ldots, u_{n^*})$ where for a fixed $j \in [n]$, we define $j^* = \arg\max_k \|u_k - u_j\|_2^2$.

  - For a full attention mechanism with appropriate query and keys, the task above is very easy as we can evaluate all pair-wise inner products

Proposition 1. *There exists a single layer full self-attention $g \in \mathcal{T}^{H=1,m=2d,q=0}$ that can evaluate Task 1, i.e. $g(u_1, ..., u_n) = [u_{1^*}, \ldots, u_{n^*}]$, but for any sparse-attention graph D with $\tilde{O}(n)$ edges (i.e. inner product evaluations), would require $\tilde{\Omega}(n^{1-o(1)})$ layers.*

Part 2: Every Sparse Attention Mechanism will need $\tilde{\Omega}(n^{1-\epsilon})$ layers We prove by contradiction that it is impossible to solve Task 1 by any $g \in \mathcal{T}_D^{H=O(d),m=O(d),q=O(d)}$ sparse-attention graph D with $\tilde{O}(n)$ edges.

Suppose we can solve Task 1 using a network $g \in \mathcal{T}_D^{H=O(d),m=O(d),q=O(d)}$ that has l layers. Recall that all the computation we do in one layer is:

$$\begin{aligned} a_i &= \textsc{Attn}_D(Q(x_i), K(X_{N(i)}), V(X_{N(i)}) + x_i \\ x_i &= O(a_i) + a_i \end{aligned} \tag{12}$$

where $\textsc{Attn}_D$ is defined in eq. (AT).

Thus, total computation per layer is $\tilde{O}(nd^3)$ and consequently $\tilde{O}(nld^3)$ for the whole network consisting of l layers.

We can use the result of Task 1 to solve the orthogonal vector (OV) problem (defined in Conjecture 1) in linear time. So in total, we will be able to solve any instance of OV in $\tilde{O}(nld^3)$ time.

Now if $l = O(n^{1-\epsilon})$ for any $\epsilon > 0$ and $d = \Theta(\log^2 n)$, then it appears that we are able to solve OV in $\tilde{O}(n^{2-\epsilon})$ which contradicts Conjecture 1. Therefore, we need at least $\tilde{\Omega}(n^{1-o(1)})$ layers. □

실험 결과

- Longer Masked LM Pretraining improves performance on downstream tasks
  - Pretraining the model

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 8: Hyperparameters for the two BIGBIRD base models for MLM.

Dataset	# tokens	Avg. doc len.
Books [110]	1.0B	37K
CC-News [34]	7.4B	561
Stories [89]	7.7B	8.2K
Wikipedia	3.1B	592

Table 9: Dataset used for pre training.

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**

Table 10: MLM performance on held-out set.

실험 결과

- Longer Masked LM Pretraining improves performance on downstream tasks
  - Question answering

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

Model	HotpotQA			NaturalQ		TriviaQA		WikiHop
	Ans	Sup	Joint	LA	SA	Full	Verified	MCQ
HGN [26]	**82.2**	88.5	**74.2**	-	-	-	-	-
GSAN	81.6	88.7	73.9	-	-	-	-	-
ReflectionNet [32]	-	-	-	77.1	**64.1**	-	-	-
RikiNet-v2 [61]	-	-	-	76.1	61.3	-	-	-
Fusion-in-Decoder [39]	-	-	-	-	-	84.4	90.3	-
SpanBERT [42]	-	-	-	-	-	79.1	86.6	-
MRC-GCN [87]	-	-	-	-	-	-	-	78.3
MultiHop [14]	-	-	-	-	-	-	-	76.5
Longformer [8]	81.2	88.3	73.2	-	-	77.3	85.3	81.9
BIGBIRD-ETC	81.2	**89.1**	73.6	**77.8**	57.9	**84.5**	**92.4**	**82.3**

Table 3: Fine-tuning results on **Test** set for QA tasks. The Test results (F1 for HotpotQA, Natural Questions, TriviaQA, and Accuracy for WikiHop) have been picked from their respective leaderboard. For each task the top-3 leaders were picked not including BIGBIRD-etc. **For Natural Questions Long Answer (LA), TriviaQA, and WikiHop, BIGBIRD-ETC is the new state-of-the-art**. On HotpotQA we are third in the leaderboard by F1 and second by Exact Match (EM).

실험 결과

- QA Dataset format

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 1: An example of the multi-hop questions in HOTPOTQA. We also highlight the supporting facts in *blue italics*, which are also part of the dataset.

Question:

who played will on as the world turns?

Short Answer:

Jesse Soffer

Long Answer:

William " Will " Harold Ryan Munson is a fictional character on the CBS soap opera As the World Turns . He was portrayed by Jesse Soffer on recurring basis from September 2004 to March 2005 , after which he got a contract as a regular . Soffer left the show on April 4 , 2008 and made a brief return in July 2010.

실험 결과

- Longer Masked LM Pretraining improves performance on downstream tasks
  - 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 [68]	29.47	6.95	26.30	37.15	11.36	33.43	27.44	7.08	23.66
	LexRank [25]	33.85	10.73	28.99	39.19	13.89	34.59	35.57	10.47	29.03
	LSA [97]	29.91	7.42	25.67	33.89	9.93	29.70	-	-	-
	Attn-Seq2Seq [85]	29.30	6.00	25.56	31.55	8.52	27.38	28.74	7.87	24.66
	Pntr-Gen-Seq2Seq [77]	32.06	9.04	25.16	35.86	10.22	29.69	33.14	11.63	28.55
	Long-Doc-Seq2Seq [20]	35.80	11.05	31.80	38.93	15.37	35.21	-	-	-
	Sent-CLF [81]	34.01	8.71	30.41	45.01	19.91	41.16	36.20	10.99	31.83
	Sent-PTR [81]	42.32	15.63	38.06	43.30	17.92	39.47	34.21	10.78	30.07
	Extr-Abst-TLM [81]	41.62	14.69	38.03	42.13	16.27	39.21	38.65	12.31	34.09
	Dancer [31]	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 [76]	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 4: Summarization ROUGE score for long documents.

실험 결과

- Longer input sequence handling capability of BIGBIRD would beneficial as many functional effects in DNA are highly non-local
  - BPC(Bits-per-character) is a metric that tells how close the LM's distribution is close to data dist.

Model	BPC
SRILM [58]	1.57
BERT (sqln. 512)	1.23
BIGBIRD (sqln. 4096)	**1.12**

Table 5: MLM BPC

Model	F1
CNNProm [90]	69.7
DeePromoter [71]	95.6
BIGBIRD	**99.9**

Table 6: Comparison.

Model	TF	HM	DHS
gkm-SVM [30]	89.6	-	-
DeepSea [109]	95.8	85.6	**92.3**
BIGBIRD	**96.1**	**88.7**	92.1

Table 7: Chromatin-Profile Prediction

실험 결과

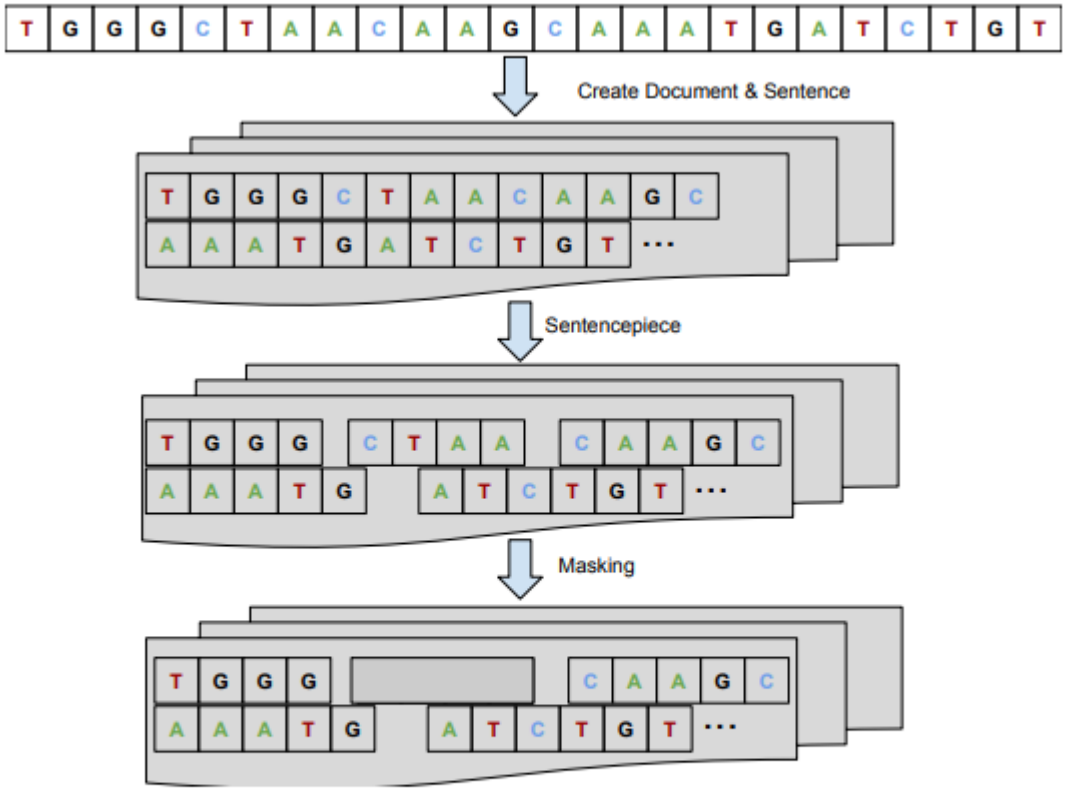


Figure 7: Visual description of how the masked language modeling data was generated from raw DNA dataset. The raw DNA sequences of GRCh37, where split at random positions to create documents with 50-100 sentences where each sentence was 500-1000 base pairs (bps). Thus each document had a continuous strand of 25000-100,000 bps of DNA. This process was repeated 10 times to create 10 sets of document for each chromosome of GRCH37. The resulting set of documents was then passed through Sentencepiece that created tokens of average 8bp. For pretraining we used masked language model and masked 10% of the tokens and trained on predicting the masked tokens.

결론

- BIGBIRD uses a sparse attention mechanism that is linear in the number of tokens
- BIGBIRD is a universal approximator of sequence to sequence functions
- BIGBIRD is Turing Complete
- Key point behind the sparse attention mechanism that authors proposed was the power of extra global tokens that preserve the expressive powers of the model
- Sparse attention mechanism do incur a cost