Big Bird: Transformers for Longer Sequences

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti,
Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, Amr Ahmed

Google Research

KIE(Knowledge and Information Engineering Lab)
배현진

1. Introduction

- Transformer 기반의 모델 → 다양한 NLP 분야에서 성과 & 최근 NLP 연구의 중심
- Self-Attention: 입력 시퀀스를 구성하고 있는 토큰들이 이전 시각의 영향을 받는 것이 아니라 그 시퀀스 내 다른 토큰에 독립적으로 attend 할 수 있도록 만들어주는 기법
  - Recurrent neural network의 sequential dependency를 없애고 병렬적으로 input sequence의 각 토큰을 처리
  - 계산 속도를 높여 엄청나게 큰 사이즈의 데이터셋으로 모델을 학습할 수 있도록 만들어 줌
  - 이런 큰 스케일의 데이터를 학습함으로써 BERT나 T5가 등장할 수 있었음!
    - General purpose corpora로 pretrain 한 다음 down-stream task에 transfer
- 하지만 full self-attention mechanism => 시퀀스 길이의 제곱에 비례하는 계산(quadratic in the sequence length)과 메모리 complexity 가짐
  - 이런 계산 부담을 덜기 위해 Attention 기법의 개선에 대한 연구가 이뤄지고 있음

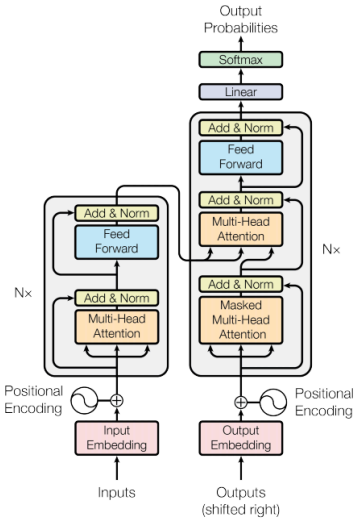
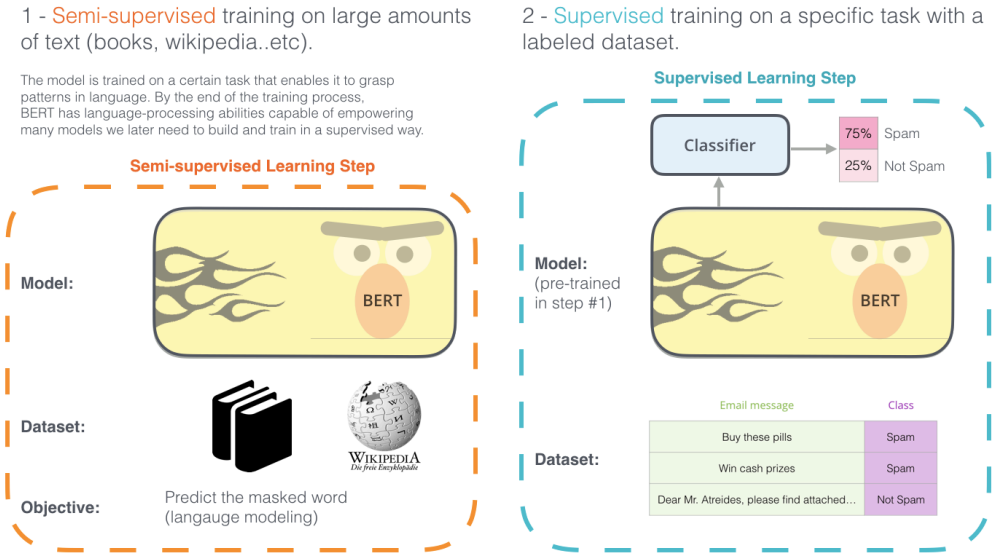


Figure 1: The Transformer - model architecture.

Transformer 구조



Downstream Task 예시

1. Introduction

- 긴 시퀀스에 대한 성능을 높이고, 이런 계산량을 줄이고자 Sparse Attention mechanism을 제시
  - BigBird: 입력 시퀀스 토큰 길이에 linear한 complexity를 가지는 Attention 매커니즘
  - 3가지 종류의 attention으로 구성
    - global attention: 시퀀스의 모든 부분에 attend
    - random attention: 각 쿼리마다 r개의 random 키에 attend
    - window attention: 로컬 neighbors w개에 attend
- BigBird가 기존의 full attention mechanism(Transformer)의 특성을 만족시킴을 증명
  - Universal approximators of Seq2Seq
  - Turing complete

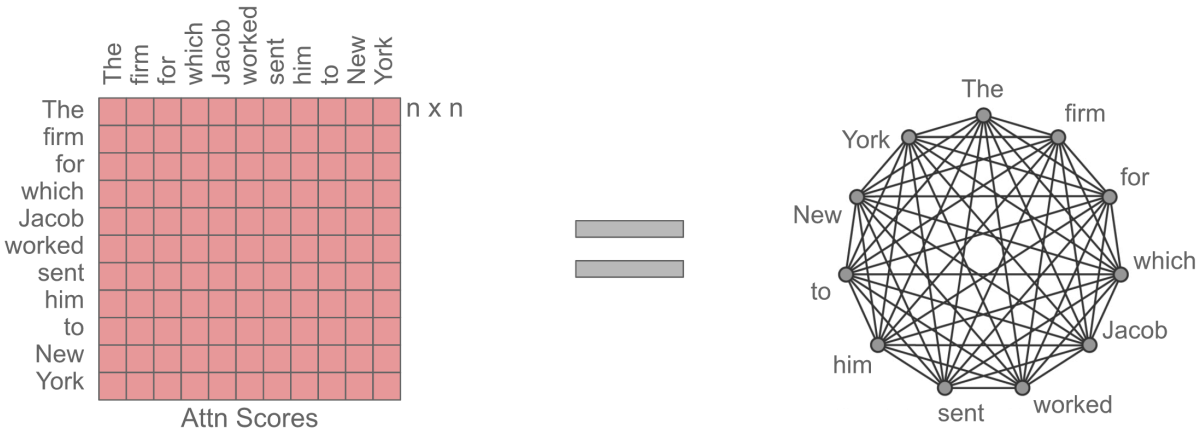
2. Architecture

- Full attention에 대한 complexity를 줄이는 것을 graph sparsification problem으로 바라볼 수 있다.
- Input Sequence $X = (x_1, \dots, x_n) \in \mathbb{R}^{n \times d}$에 대해 generalized attention mechanism을 적용:

  → vertex set $[n] = \{1, \dots, n\}$에 대한 directed graph D

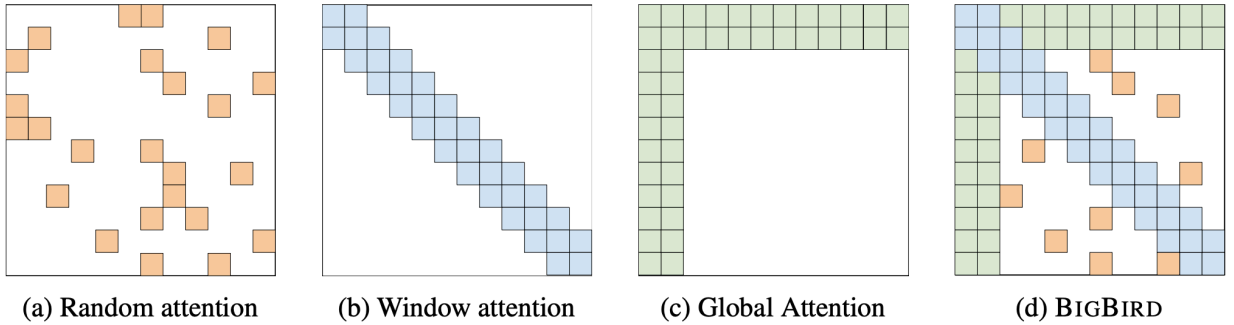
  - 각 노드들에 대한 directed edge로 그 노드에 대한 어텐션을 표시
- 그래프 D가 모든 노드가 연결되어 있는 complete bigraph이면 기존의 full attention mechanism

  - n: sequence length
  - d: embedding dimension



2. Architecture

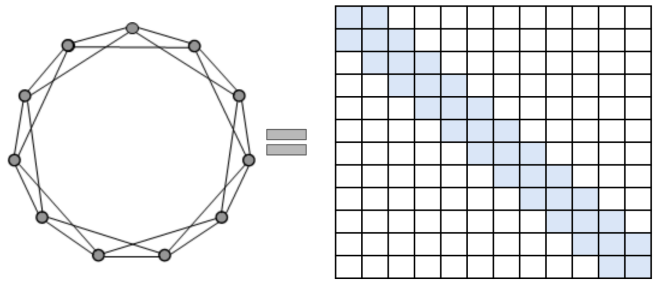
- 3가지 attention을 합친 구조



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

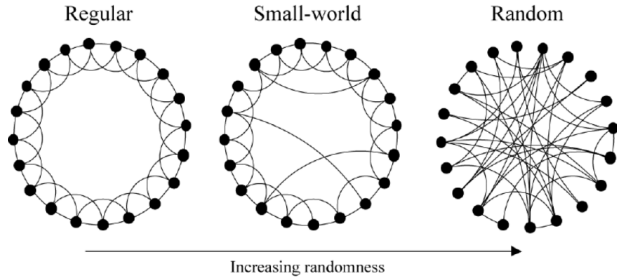
Figure 1: Building blocks of the attention mechanism used in BIGBIRD. White color indicates absence of attention. (a) random attention with $r = 2$, (b) sliding window attention with $w = 3$ (c) global attention with $g = 2$. (d) the combined BIGBIRD model.

Window attention(Locality)

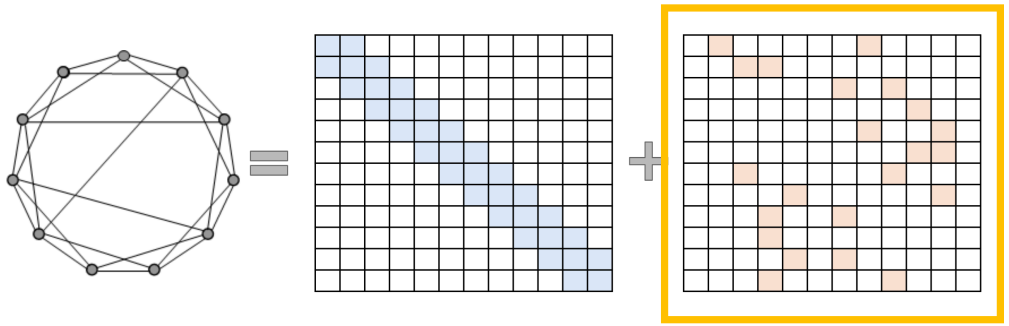


- 언어적 구조에 기반, 많은 경우에 해당 토큰에 대한 정보는 이웃한 토큰 사이에서 찾을 수 있음
- Clark et al(2019): 자연어 처리에 사용되는 self-attention 모델에서 인접한 inner-products가 가장 큰 영향을 미친다.
- sliding window attention을 적용
  - 현 위치 i에서 $(\mathrm{i} - \frac{w}{2}) : (i + \frac{w}{2})$개의 key 고려

2. Architecture

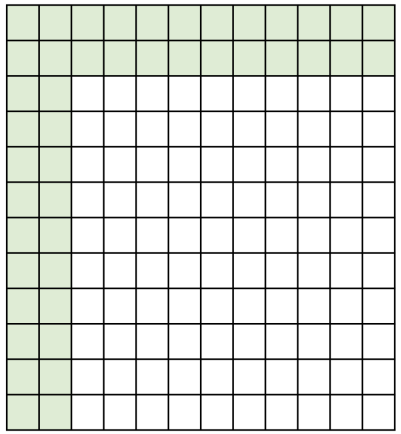


Random Attention



- Erd ős-Rényi model: 가장 간단한 랜덤 그래프 모델 중 하나로 고정된 확률값에 의해 각 간선이 독립적으로 정해지는 모델
  - 이러한 랜덤 특성은 노드 간 정보가 빨리 전달될 수 있도록 만듦
  - 결과적으로 이런 랜덤 그래프는 complete graph에 근사하게 됨
- 이를 이용해 각 query가 랜덤한 r개의 key에 attend하는 sparse attention을 제안

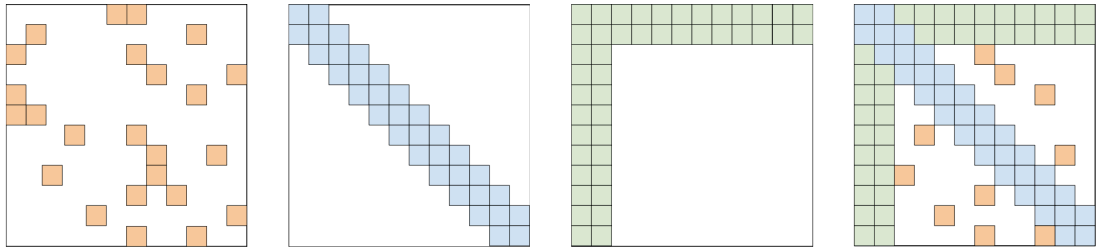
Global Attention



- Random & Window Attention을 적용해 실험한 결과 Full attention 모델 성능에 미치지 못하는 결과 보임
- 따라서 모든 token에 집중 하면서 모든 token으로부터 attention을 받는 global token 도입
  - BigBird-ITC: 기존의 토큰 중 G개의 subset을 global token으로 만듦
  - BigBird-ETC: BERT의 'CLS'와 같은 새로운 global token을 추가

3. Theoretical Results about Sparse Attention Mechanism

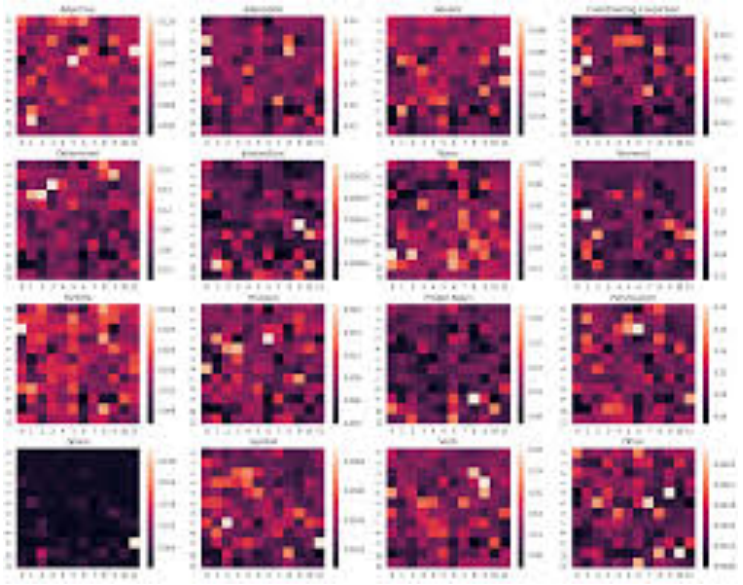
- BigBird: 앞서 설명한 3가지 Attention을 합한 모델
  - 밑의 그림과 같은 Full attention 모델과 비교하면, Sparse한 형태를 보임



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

- 그렇다면, 이런 sparse 한 모델이 Full attention의 특성을 다 만족할까?
- 두 가지 측면에서 full attention 특성 모두 만족시킨다는 점을 증명

1. Seq2Seq의 Universal Approximator
   - Universal Approximation Theorem: 어떤 함수로 근사시킬 수 있다는 이론
   - Yun et al.(2019): Transformer가 Seq2Seq의 Universal Approximator임을 증명
2. Turing Complete
   - Pérez et al.(2019): Transformer에 대한 튜링 완전성 증명



4. Experiments

- Pretrained Masked LM
- Encoder only task
  - QA: 긴 sequence에서 answer에 대한 evidence를 찾을 수 있는지
  - Document Classification: 긴 Document 내에서 정보를 찾을 수 있는지
- Encoder-Decoder task
  - summarization
- Genomics

4. Experiments

Pretrained Masked LM

- BigBird base & large을 MLM을 이용해 pretrain
  - 마스킹 된 부분에 대한 토큰들을 예측

Dataset	# tokens	Avg. doc len.
Books [110]	1.0B	37K
CC-News [35]	7.4B	561
Stories [90]	7.7B	8.2K
Wikipedia	3.1B	592

Table 2: 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 3: MLM BPC on held-out set.

- Longformer: 48GB 메모리 + batch size of 16-32
- Ours: 16GB 메모리 + batch size of 32-64

4. Experiments

Encoder Only Tasks: QA

- NaturalQ
- HotspotQA
- TriviaQA-wiki
- WikiHop

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.

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**

Table 5: 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 Verified, 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).

4. Experiments

Encoder Only Tasks: Document Classification

- 적은 training data과 긴 input data를 가질 수록 improvement가 높았음

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.

4. Experiments

Encoder-Decoder Tasks - Abstractive 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.

5. Conclusions

- Sparse Attention mechanism **BigBird** 고안
  - 기존 BERT보다 8배 긴 시퀀스 처리 가능
  - $O(n^2) \rightarrow O(n)$
- 해당 Sparse Attention이 Full Attention의 특성 두가지를 만족함을 증명
- 다양한 task에서 sota 달성