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

* RNN의 sequential dependency를 없애고 병렬적으로 input sequence의 각 토큰을 처리
* 계산 속도를 높여 엄청나게 큰 사이즈의 데이터셋으로 모델을 학습할 수 있도록 만들어 줌

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

긴 시퀀스에 대한 성능을 높이고, 이런 계산량을 줄이고자 Sparse Attention mechanism을 제시

BigBird: Full-Attention 장점은 그대로, 연산량을 줄일 방법을 고안하며 만든 모델

3가지 종류의 attention으로 구성

* 랜덤 어텐션: 각 쿼리가 랜덤하게 키를 참고하도록 하는 방법
* 윈도우 어텐션: 비슷한 의미의 단어들이 비슷한 위치에 있다는 것을 감안, 각 쿼리가 양옆에 있는 키만 참고하도록 하는 방법
* 글로벌 어텐션: 일부 쿼리의 경우 모든 키를 참고하도록 하는 방법

기존 BERT 모델보다 8배 긴 시퀀스 처리 가능

시간복잡도: O(n^2) -> O(n)

KoBigBird는 이러한 BigBird 모델을 이용해 한국어 처리를 더 잘 할 수 있도록 고안된 모델