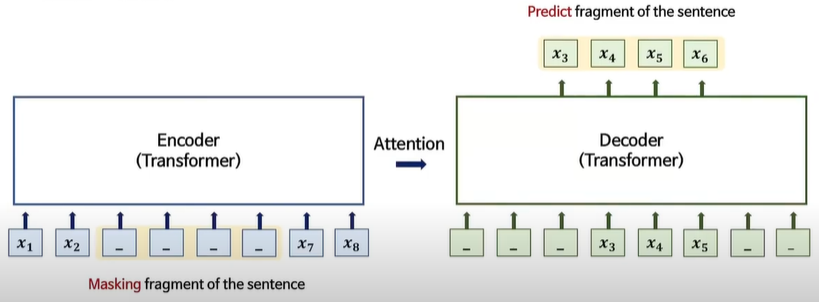
[**https://www.youtube.com/watch?v=v7diENO2mEA&feature=youtu.be**](https://www.youtube.com/watch?v=v7diENO2mEA&feature=youtu.be)

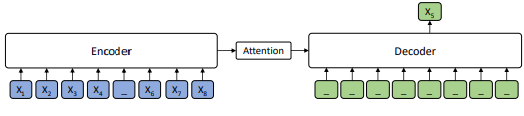
**MASS: Masked Sequence to Sequence Pre-Training for Language Generation  
(ICML 2019)**

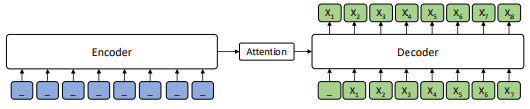


인코더와 디코더의 입력이 다름 -> 서로 상호 의존적으로 학습하게 된다.

그 이유는 디코더의 처음은 <s> 토큰이라 할 때, 인코더에서 마스킹하지 않은 토큰은 디코더에서는 마스킹 처리 돼 들어가게 된다.

이때, k-length fragment 만큼 Masking 하게 되는데 k는 하이퍼 파라미터이다.





만약 k=1 이라면 기존 BERT와 입출력 단의 형태는 같다.(단지 인코더-디코더 구조를 여기서 이용)

만약 k=m 이라면 기존 GPT와 입출력 단의 형태는 같다.(단지 인코더-디코더 구조를 여기서 이용)

이게 왜 좋나..?

+ 결과는 나중에