<https://blog.pingpong.us/electra-review/>

**ELECTRA: Pre-Training Text Encoders as Discriminators Rather Than Generators(ICLR 2020)  
(Efficiently Learning an Encoder that Classifies Token Replacements Accurately)**

이 논문은 충분한 컴퓨팅 리소스가 없는 영세 연구자들은 학습하기 어렵우며 기존 상당한 계산량을 지적하며 학습의 효율성 + 모델의 정확도를 올리는 RTD라는 pre-training Task 제안

모델 크기, 데이터, 컴퓨팅 리소스가 동일한 조건일 때, 기존 BERT의 성능 능가

하나의 GPU로 4일 학습했는데 GPT를 능가

RoBERTa, XLNet의 1/4 계산량으로 비슷한 성능

(최종적으로는 성능이 똑같은데 빠르게 학습하니 그 점을 강조하는 듯 싶다.)

**Introduction**

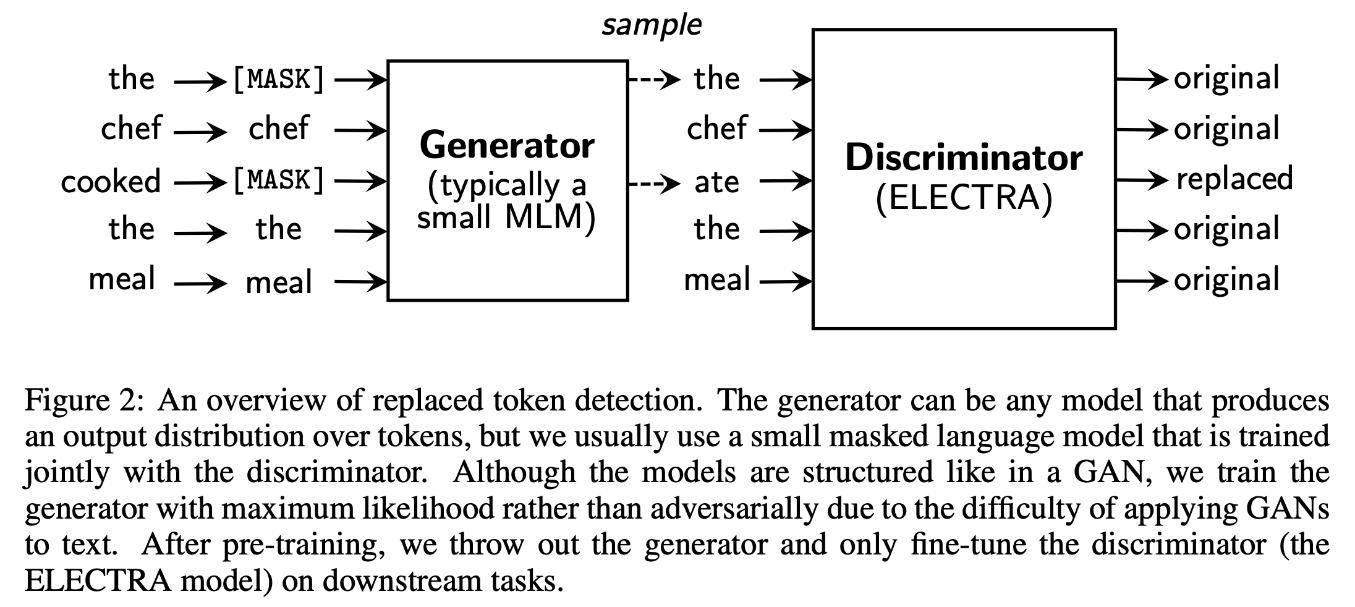
MLM의 denoising autoencoder 학습은 SOTA에 많이 사용하지만 단점이 있다.

1. 전체 토큰 중 15%에 대해서만 MASK 작업이 있고 그에 대해서만 loss가 발생  
   -> 그것만 학습 나머지 85%는 장식품
2. 따라서 학습하는 데 더 많은 step -> 즉 시간이 많이 듦
3. 학습 때는 MASK 토큰을 모델이 참고하지만 inference 때는 MASK 토큰이 존재하지 않음 (현실과의 bias)

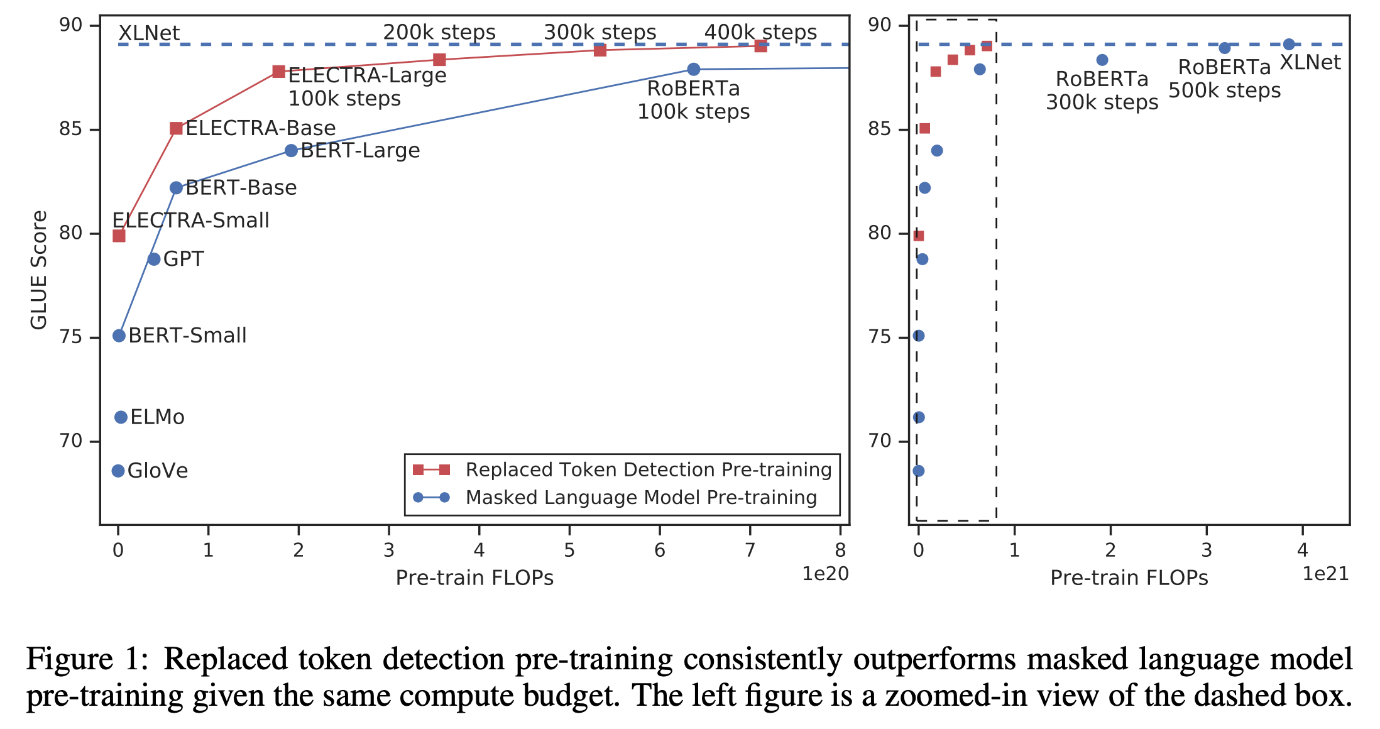
따라서 Replaced Token Detection (RTD) 라는 새로운 pre-training Task를 제안

Generator: 실제 입력의 일부 토큰을 그럴싸한 가짜 토큰으로 바꿈

Discriminator: 각 토큰이 original or replaced 인지 맞추는 이진 분류!



따라서 RTD는 Task로 입력의 15%가 아닌 모든 토큰에 대해서 학습하기 때문에 상당히 효율적이며 효과적! (훨씬 빠르게 학습 & downstream task에서 성능 Up)



빠르고 성능이 향상됨, 동일한 조건에서는 모든 학습 과정에서 더 높은 GLUE 성능!

ELECTRA-small은 하나의 GPU로 4일이면 학습 가능(BERT-Large의 파라미터 수 1/20, 계산량 1/135), 근데 BERT-Small 보다 5 point GLUE 성능 Up, GPT1 보다도 높음

ELECTRA-large는 RoBERTa, XLNet 보다 더 적은 파라미터, 1/4 계산량이지만 비슷한 성능, GLUE에서는 ALBERT 능가, SQuAD 2.0에서는 SOTA -> 효율적으로 학습