3.2. RoBERTa : A Robustly Optimized BERT Pretraining Approach

논문 : https://arxiv.org/abs/1907.11692

5.2. XLNet 이후에 나온 논문

BERT의 Pretrain 과정을 수정한 논문

RoBERTa는 BERT 모델을 Pretrain을 해야될 때 참고할만한 내용

RoBERTa의 제안한 사전학습 방법

1. **더 많은 데이터 사용**
    - BERT의 Pretrain과정에서 16GB 텍스트 데이터를 활용함
    - XLNet 성능향상의 원인 중에 Pretrain에 사용한 데이터 크기가 영향이 있는 것을 확인
    - RoBERTa 도 학습데이터를 160GB로 증가함으로 BERT의 성능을 향상시킴
2. **Dynamic Masking**
    - BERT는 Pretrain과정에서 하나의 문장에 임의의 마스크 토큰을 씌워서 반복적으로 학습시킴
    - 마스크 토큰에 대한 Bias가 생기게 되므로 문제가 있음
    - RoBERTa는 40 Epochs 마다 같은 데이터에 대해서 10번의 다른 Masking을 적용해 학습시킴
3. **다음 문장 예측 (Next Sentence Prediction) 제거**
    - BERT는 NSP를 training objective에 포함시켜서 훈련을 진행
    - 후속 논문(XLNet)들은 NSP에 대해 부정적인 효과를 가져온다고 판단
    - NSP 자체가 Pretrain과정에서 하지않아도 fine-tuning과정에서 충분히 학습할 수 있다고 판단되서 제거(NSP 자체가 너무 쉬운 문제)
4. **배치 사이즈 증가**
    - 기존의 언어 모델은 작은 배치 사이즈 (한번의 스탭에 256개)로 학습을 진행
    - BERT 모델은 더 큰 배치사이즈를 써야 잘 훈련되는 것을 확인
5. **Word Piece 방식에서 Byte Pair Encoding으로 변환**
    - BERT는 유니코드 글자 단위의 Word Piece Vocabulary를 사용
    - 이 경우 Out-of-Vocabulary가 발생해 언어 모델의 성능을 저하시키는 요인이 되버림
    - GPT 처럼 Byte Pair Encoding으로 더 많은 글자를 다룰 수 있게 변환