CS224N 발표

2025.05.12

발표자 : 김지호

Introduction

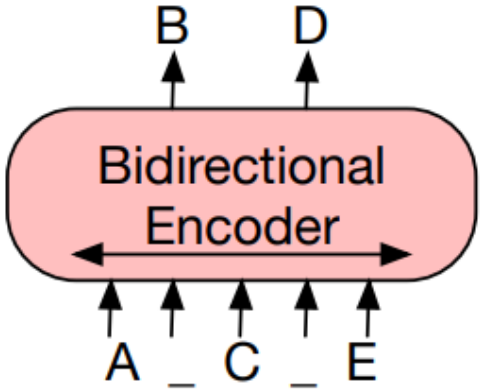
BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension

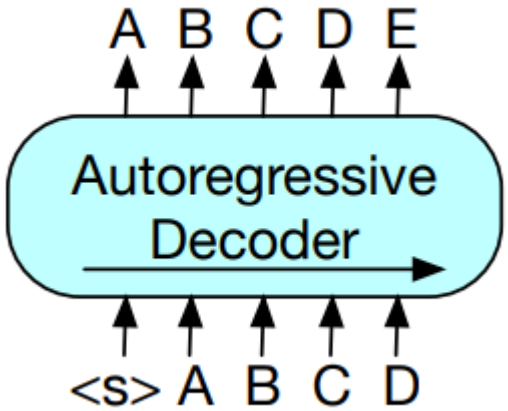
2019년 Facebook AI 팀에서 발표

Introduction

기존 BERT, GPT의 문제점

BERT: Generation Task에 사용하기 어렵다는 문제
GPT: 생성은 잘 하나 양방향 정보를 파악할 수 없다는 문제



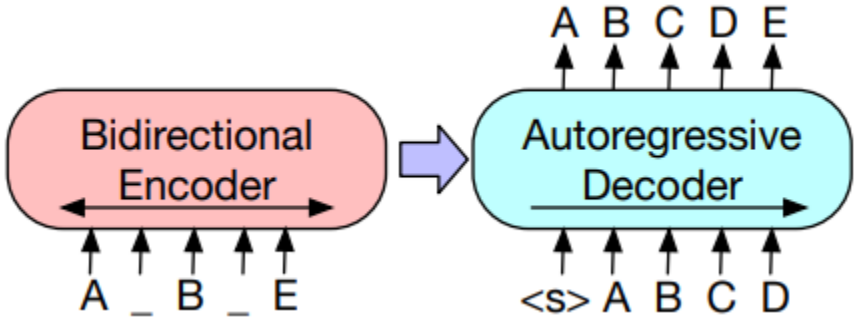


Introduction

Bidirectional-Autoregressive Transformers

BART: BERT의 bidirectional + GPT의 autoregressive

Denoising Autoencoder: 텍스트를 noise로 손상시키고, 이를 다시 원본 텍스트로 복원하도록 학습함



Model

BART의 전체적인 모델 구조는 Transformer의 구조를 따르나, ReLU 대신 GeLU 사용
매개변수를 N(0, 0.02) 분포로 초기화

GeLU: Gaussian Error Linear Unit

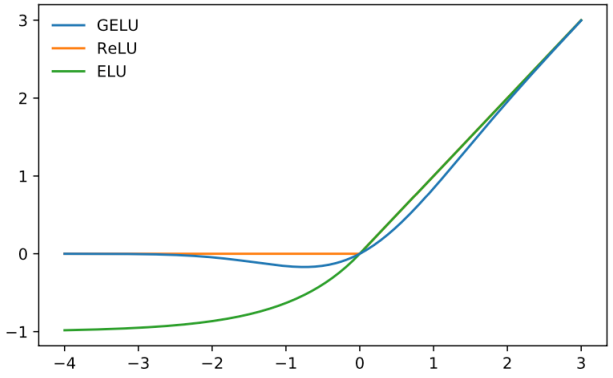


Figure 1: The GELU ($\mu = 0, \sigma = 1$), ReLU, and ELU ($\alpha = 1$).

$$\mathrm{GELU}(x) = xP(X \leq x) = x\Phi(x)$$

표준정규분포의 CDF를 이용
X가 크면 P(X <= x) = 1
X가 0에 가까우면 P(X <= x) = 0.5
X가 작으면 P(X <= x) = 0

Transformer류 모델에서 ReLU에 비해 좋은 성능을 보이나 연산량이 증가한다는 단점도 있음

Model

BART-base: 인코더, 디코더 각각 6개 층 (총 12층)
BART-large: 인코더, 디코더 각각 12개 층 (총 24층)

BERT의 구조와 유사하지만
디코더의 각 층이 인코더의 마지막 은닉층에 대해 Cross-attention 연산 수행

BERT의 경우 단어 예측 전 추가적인 feed-forward network를
사용하지만 BART는 사용하지 않는다.

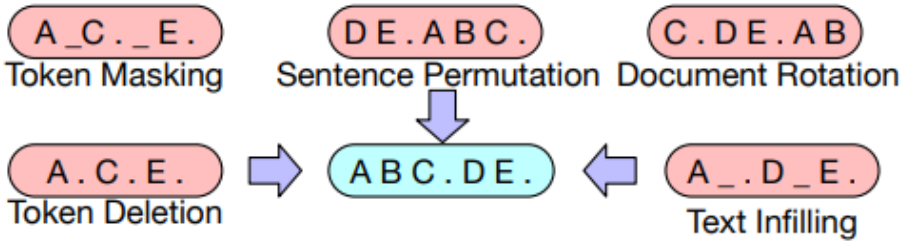
Pre-training BART

BART의 학습: 문서 훼손을 통해 오염된 값을 다시 복원하여 디코더의 출력과 원래 입력값의 Reconsturction loss(Cross entropy)를 최적화하여 학습

학습을 위한 BART의 다양한 문서 noising 기법

1. Token Masking: BERT와 같이 무작위 토큰을 [MASK]로 대체

2. Token Deletion: 입력값에서 무작위 토큰을 삭제. 모델이 누락된 위치를 추론



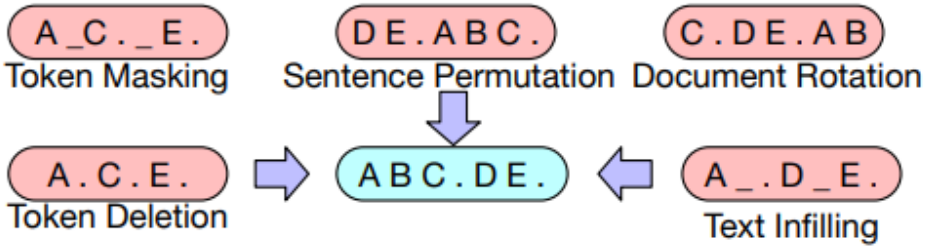
Pre-training BART

BART의 다양한 문서 noising 방법

3. Text Infilling: Poisson(3) 분포에서 선택된 랜덤한 길이의 Span을 단일한 [MASK]로 대체. 모델은 Span에서 누락된 토큰의 수를 추정

4. Sentence Permutation: 문장 순서를 무작위로 섞음

5. Document Rotaion: 한 토큰을 무작위로 선택하고, 해당 토큰을 시작으로 문서가 시작되도록 회전함. 모델은 문서의 시작 부분을 식별하도록 훈련



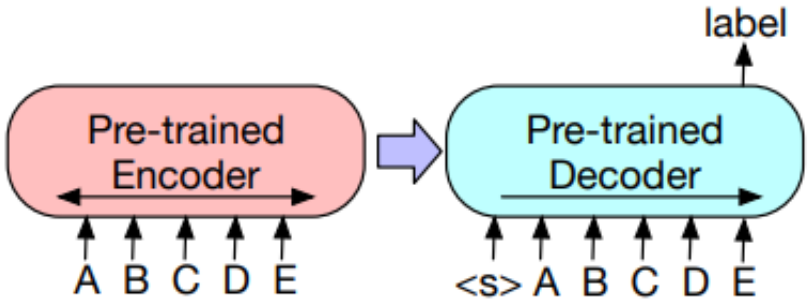
Fine-tunning BART

Sequence Classification Task

인코더와 디코더에 동일한 입력이 주어짐

마지막 디코더 토큰의 final hidden state가 multi-class linear classifier에 전달하여 classification 수행

BERT의 CLS 토큰과 유사하지만 추가적인 토큰을 넣어 전체 입력에 대한 정보를 활용



Fine-tunning BART

Sequence Classification Task

BERT: 문장 맨 앞에 [CLS] 토큰을 추가하고, 그 토큰의 벡터를 Classification에 사용

BART: 인코더, 디코더에 같은 문장을 입력. 문장 마지막에 특별 토큰을 추가

디코더는 인코더와 cross attention 수행 및 self attention을 수행하므로 마지막 특별 토큰은 문장 전체의 맥락을 담고 있음

이 토큰의 벡터를 Classifier에 넣으면 인코더+디코더 맥락을 모두 반영하여 Classification을 수행할 수 있음

Fine-tunning BART

Token Classification Task

인코더와 디코더에 전체 문서 입력
디코더의 각 최상층 hidden state를 각 단어의 representation으로 사용하여 token을 분류

Sequence Generation

인코더의 입력은 input sequence를 받고 디코더는 출력을 auto-regressive하게 생성
디코더가 이전까지 생성한 토큰 + 인코더의 context 정보를 통해 다음 토큰 예측
Autoregressive decoder를 통해 generation task에 파인튜닝이 가능

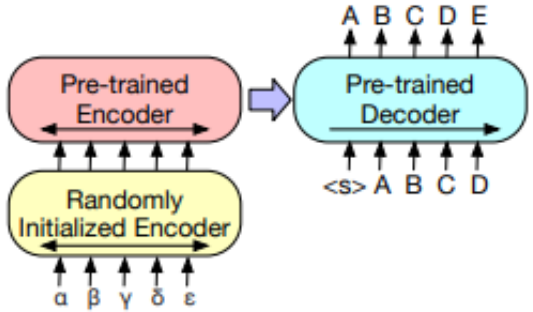
Fine-tunning BART

Machine Translation

BART 전체를 pre-trained decoder로 사용

새로운 encoder parameter를 추가하여 외국어를 BART가 처리 가능한 형태로 변환

Randomly initialized encoder는 외국어 토큰을 노이즈가 낀 잠재적 영어 표현으로 바꿈



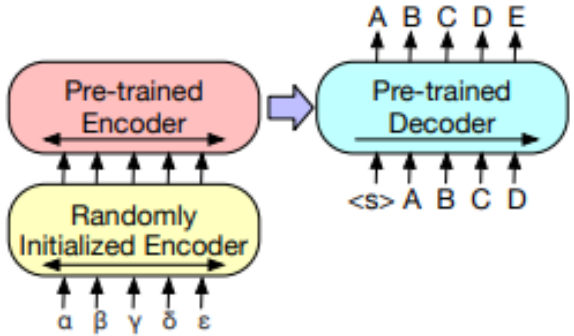
Fine-tunning BART

Machine Translation

이 시퀀스 벡터가 BART의 인코더로 들어가면 정답 영어 단어에 맞게 파인튜닝됨

첫번째 단계에서는 BART 인코더의 거의 모든 parameter를 freeze하고, source encoder만 업데이트

다음 단계에서는 BART 전체 모델 파라미터를 학습



Fine-tunning BART

Machine Traslation

BART의 인코더+디코더는 마치 노이즈를 제거하듯 영어 문장 생성

즉 Randomly initialized encoder: 외국어->잠재 영어 표현

BART(인코더+디코더): 잠재 영어 표현 -> 올바른 영어 표현
역할을 수행함