BERT

Pre-training of Deep Bidirectional Transformers for Language Understanding

25.05.07

정보컴퓨터공학부
202055528 김태환

Contents

소개

- **BERT 란 ?**
  - language presentation Model 이다.

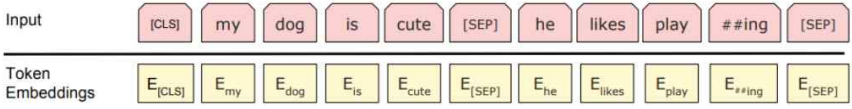
- **BERT 의 특징은 ?**
  - 타 language model 들과 달리 문장을 양방향 분석한다.
  - 단 한 개의 Layer 를 추가함으로써 광범위한 작업에 응용시킬 수 있다.
  - **Masked Language Model** 사전학습목표를 활용한다.
  - **Next Sentence Prediction** 을 활용한다.

- **BERT 의 퍼포먼스 ?**
  - 개념적으로 매우 간단하지만, 강력한 성능을 보인다.

모델구조

- **Masked Language Model 이란 ?**
  - 각 문장 내의 token 들은 고유의 ID 를 가진다.
  - 무작위로 문장 내의 token 을 Masking 한다.
  - Mask 의 좌우 문맥을 융합시킴으로써 양방향 모델로 학습된다.

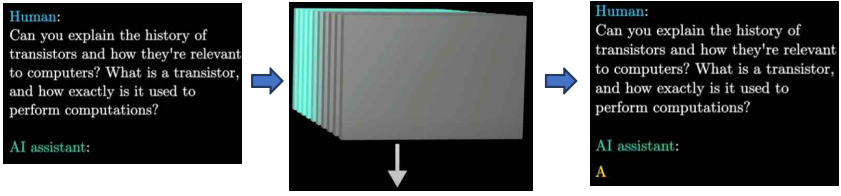
- **Next Sentence Prediction 이란 ?**
  - token 간의 관계가 아닌, sentence 간의 관계를 학습시킨다.



모델구조

- **BERT 의 구현 소개**
  - **두 개의 step**
    - **pre-training** : model 을 unlabeled data 로 학습시킨다. (하위에 MLM, NSP)
    - **fine-tuning** : 실제로 해결하고자 하는 task 를 위해 모델을 추가학습시킨다.

  - **Transformer 의 Encoder 모델을 차용한다**
    - BERT 의 구조는 Multi-layer bi-directional transformer encoder 이다.
    - GPT 는 단방향 self-attention transformer 구조, BERT 는 양방향 self-attention transformer 구조이다.

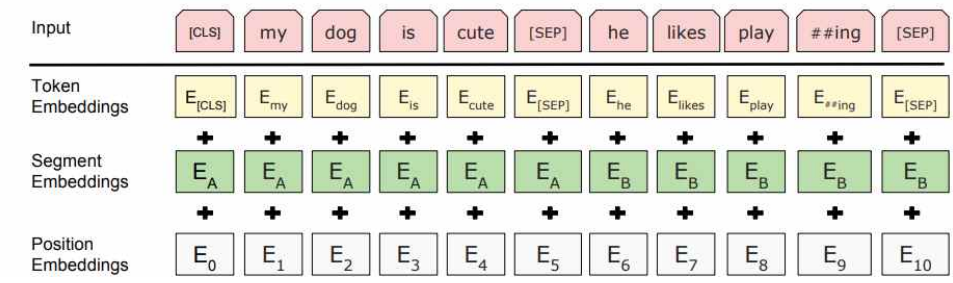


https://www.youtube.com/watch?v=LPZh9BOjkQs

모델구조

- **BERT 의 입출력 소개**
  - **각 token 은 Word Piece 를 통해 embedding 이 진행되어 입력된다.**
  - **Sequence 는 BERT 에 들어가는 input 이다.**
    - Sequence 는 하나의 sentence 이거나, 두 개의 sentences 이다.
  - **입력 sequence 의 첫번째 토큰은 [CLS] 이다.**
    - 해당 token 의 출력 vector 는 sequence 의 분류 정보를 담고있다.
  - **sentence pairs 는 하나의 sequence 로 묶인다.**
    - 각 sentence 는 [SEP] 토큰으로 나누어지거나, 토큰이 속한 sentence 정보를 벡터에 반영한다.
  - **각각의 token 의 입력 vector 는 세 개의 embedding 정보가 합쳐진 값이다.**
    - token 의 embedding 값 + segment 정보(sentence) + position 정보(sentence 내 위치)

모델구조



- token 의 embedding 값 + setgment 정보(sentence) + position 정보(sentence 내 위치)

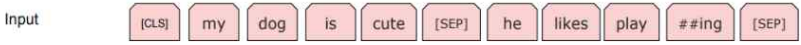
모델구조

- **Pre-training**
  - **Model 의 문맥 이해 능력을 키우키 위한 작업이다.**
  - **BERT 는 두 개의 방법으로 양방향 학습을 진행한다.**
    - Masked Language Model
    - Next Sentence Prediction
  - **Masked LM**
    - 입력토큰의 일부(15%)를 가지고, 모델이 예측하도록 한다.
    - mask token 으로부터 출력되는 vector 는 모든 단어들에 대한 output softmax 로 들어간다. (단어들에 대한 확률 계산)
    - **하지만 실제 모델을 사용할 때, mask 는 존재하지 않는다.**
      - mask 로 선택된 token 을 [MASK], random token, unchanged token 으로 각 80%, 10%, 10% 확률로 바꾼다.
      - 위의 방법으로 훈련과 활용과정에서 생기는 차이를 막는다.
    - mask 로 선택된 token 의 출력 vector 는 원래의 token 값을 예측하는데 활용된다.

왜 비지도학습이라고 하는가? => 사람이 직접 데이터를 라벨링하지 않고, 데이터 자체를 모델에게 던져주기에!

모델구조

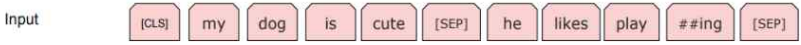
- **Next Sentence Prediction**
  - 질문-응답 작업이나, 자연어추론 작업의 경우 여러 sentence 간의 관계를 이해하는데에 기반한다.
  - 모델로 하여금 sentence 뒤에 오는 sentence 가 타당한지 그렇지 않은지 맞추도록 학습시킨다.
  - sentence A 뒤에 오는 sentence B 를 실제로 A 다음으로 올 sentence, 무작위로 뽑은 sentence 를 50% 확률로 배치시켜 true인지 false 인지 맞추도록 학습한다.
  - [CLS] 의 출력 vector 를 Next Sentence Prediction 에 사용한다.
    - 여러 sentence 묶음의 정보를 담고있기에, 이를 통해 연결되는 sentence 인지 아닌지 판단 가능하다.



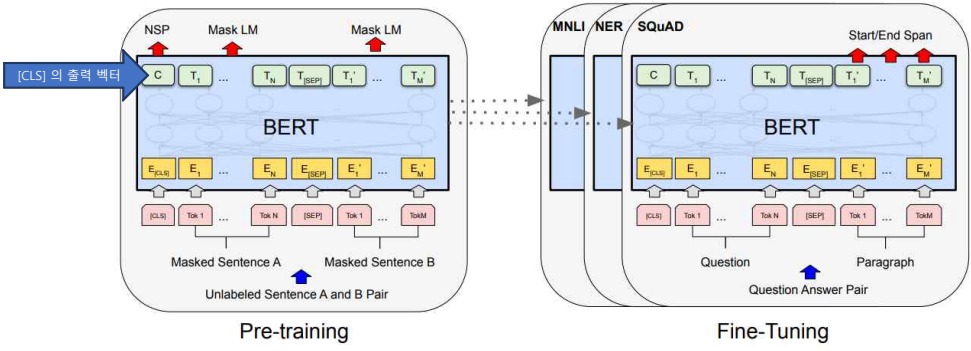
모델구조

- **Fine-tuning**
  - 기존의 text pair 를 다루는 방법은 각 sentence 를 따로 encoding 하고 관계를 계산했었다.
  - BERT 는 self-attention 을 통해 text pair 를 합친 input 을 넣어 바로 관계를 계산할 수 있다.
  - 각 task 별로 입출력을 설정하여 다시한번 학습시킨다.
  - pre-training 과정에 비해 훨씬 inexpensive 하다.

ex) QnA model 을 구축하고 싶다면 질문-정답 쌍의 text 를 넣어 학습시킨다.



모델구조


NSP
Mask LM
Mask LM
[CLS] 의 출력 벡터
BERT
Masked Sentence A
Masked Sentence B
Unlabeled Sentence A and B Pair
Pre-training
MNLI
NER
SQuAD
Start/End Span
BERT
Question
Paragraph
Question Answer Pair
Fine-Tuning

결론

- **BERT: Bidirectional Encoder Representations from Transformers**
  - 비지도 사전학습(Masked LM, Next Sentence Prediction)을 통해 언어 이해 능력을 학습한다.
  - 사전학습 후 fine-tuning 단계에서 질문응답, 문장분류, 감정분석 등 여러 작업에서 최고 성능을 달성했다.
  - Transformer 구조를 사용하여 병렬연산이 가능해졌다.
  - Bidirectional self-attention을 통해 양방향 문맥을 모두 활용할 수 있다.

감사합니다