DeBERTa: Decoding-enhanced BERT with Disentangled Attention(2020)

☑ 복습 ☐

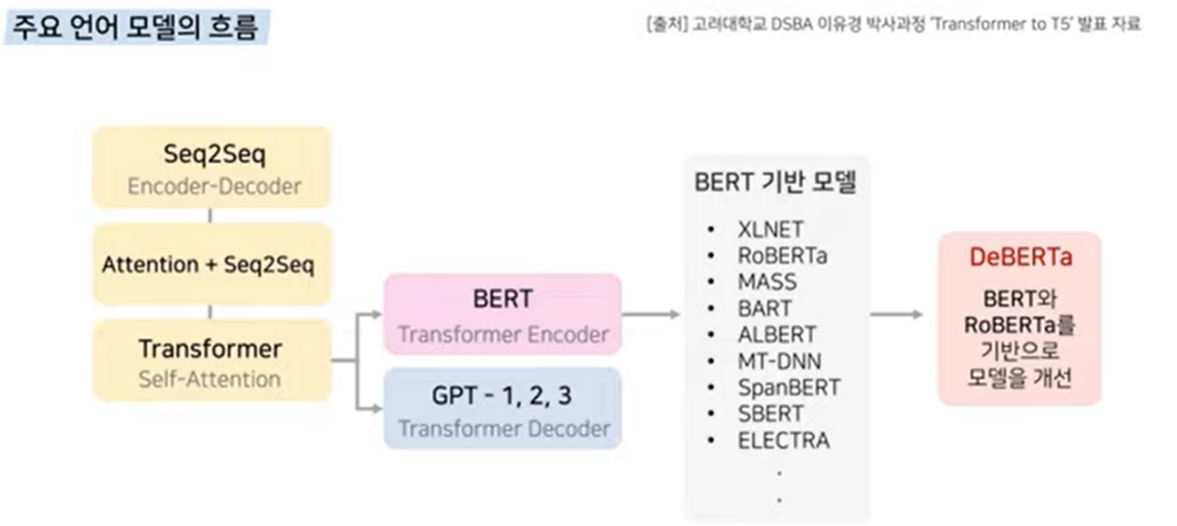
deberta_decoding_enhanced_bert.pdf

GitHub - microsoft/DeBERTa: The implementation of DeBERTa
The implementation of DeBERTa. Contribute to microsoft/DeBERTa development by creating an account on GitHub.
https://github.com/microsoft/DeBERTa

Transformer-based PLM(Pre-trained Language Model)

→ 자연어 처리 모델은 Encoder-Decoder가 결합한 형태로 변해가는 추세이며, 그 중에서도 transformer 기반의 모델이 강력한 성능을 보임

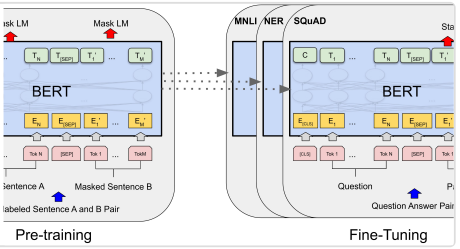
→ Transformer의 Encoder를 사용한 BERT의 Masked Language Model을 기반으로 다양한 응용 모델이 연구됨에 따라 input noise에 강건하고 양방향의 context 정보를 보다 깊게 고려 가능



BERT 언어모델

BERT(Bidirectional Encoder Representations from Transformers)는 구글에서 2018년에 발표한 언어 모델로, Transformer 아키텍처를 기반으로 하고 양방향(bidirectional) 학습을

https://velog.io/@tm011899/BERT-언어모델



`Disentangled Representation`

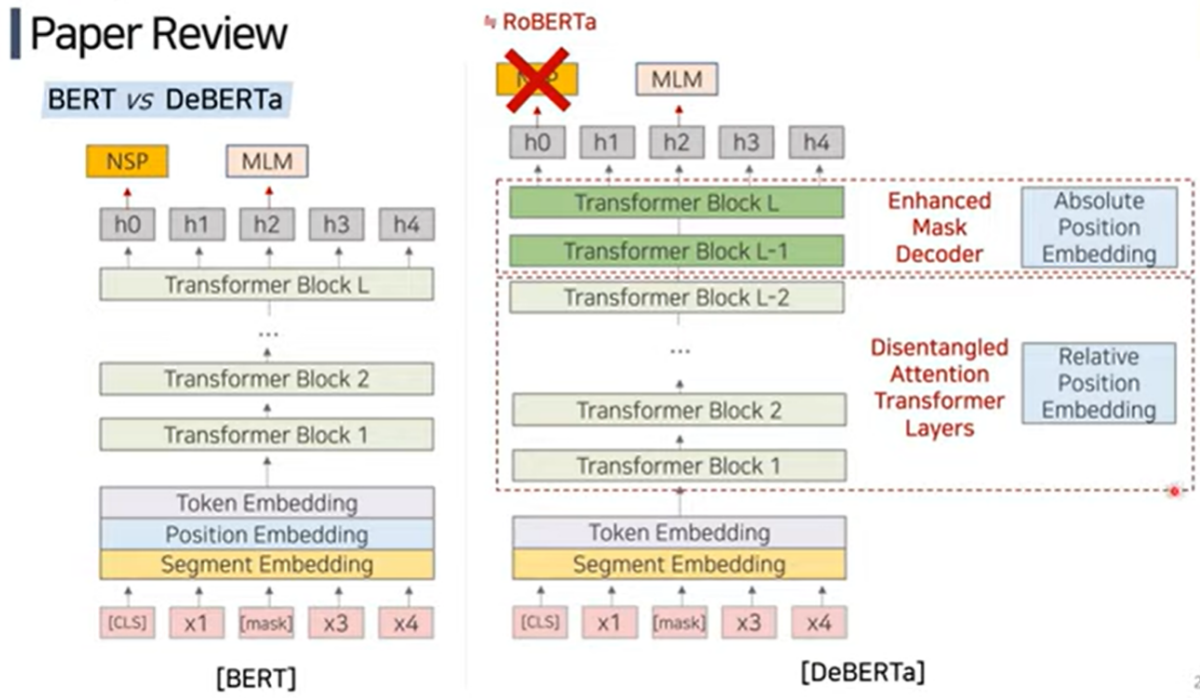
→ 서로 뒤얽혀 있는 특징 요소들을 독립적으로 풀어서 표현하기!
데이터의 다양성을 설명하는 latent 요소들을 분리하여 표현함으로써 interpretability를 높임

`Positional Embedding`

→ 단어의 위치 정보를 특정 차원의 벡터로 표현한 것

1) Absolute Positional Embedding
어떤 입력 문장이더라도 각 단어의 위치의 Positional Embedding 값은 동일한 값이 사용됨
각 토큰의 절대적 위치 정보, 절대적으로 떨어진 거리 정보를 파악할 수 있다 → distance

2) Relative Positional Embedding
특정 위치의 단어를 기준으로 일정 길이 이내에 위치한 언어와의 상대적 거리 관계 정보를 반영함
각 단어 간의 위치 차이가 index로 사용되며 짝지어진 토큰 간의 상대적 위치 정보를 파악할 수 있다, → directionn and distance



https://www.youtube.com/watch?v=hNTkpNk7v-I

1. 논문 내 주요 방법론

- Disentangled Attention Mechanism: 각 단어를 독립적인 두 벡터로 표현, 하나는 '내용(content)'을 다른 하나는 '위치(position)'을 나타낸다. 단어 간의 attention 가중치는 그들의 내용과 상대적 위치에 대해 분리된 행렬을 사용하여 계산된다.
- Enhanced Mask Decoder: 모델의 학습(pre-training) 단계에서 각 토큰의 절대적 위치(absolute position)를 디코딩 레이어에 포함시킨다.
  → 문장에서 단어가 쓰인 순서와 위치에 따른 미묘한 의미 차이를 더 잘 이해

2. 논문에서 풀고자 하는 문제

- BERT와 같은 모델들이 단어 간의 복잡한 관계를 포착하는 데 있어서 갖는 한계점
  → 단어의 정확한 위치 정보와 이들 간의 상호작용을 효과적으로 모델링하지 못했기 때문
- 기존 모델들의 단어의 내용과 위치 정보를 통합적으로 처리하는 방식
  → 모델이 단어 간의 관계를 더욱 정교하게 학습하는 데 한계 발생

3. 기존의 연구들이 이 문제를 풀어온 방식

문장 내 의미, 단어 간의 관계를 파악하기 위해 주로 어텐션 메커니즘을 사용했고, 위치 정보는 상대적인 위치 표현 또는 고정된 위치 인코딩을 통해 간접적으로만 반영했다.

4. 기존의 연구 대비 이 논문의 강점

- `Disentangled Attention`
  대부분의 트랜스포머 기반 모델은 단어와 위치 정보를 하나의 벡터로 결합하여 어텐션 계산에 사용한다. 그러나 DeBERTa는 단어의 내용(content)과 위치(position) 정보를 분리하여 처리하는 Disentangled Attention 메커니즘을 도입하여, 모델은 단어 간의 내용적인 관련성(content-wise attention)과 각 단어의 상대적인 위치에 따른 관계(position-wise attention)를 더욱 세밀하게 파악할 수 있게 되었다.
    - Transformer-xl 처럼 addictive하게 attention을 분해
    - Shaw, Transformer-xl과 다르게 position-to-content term을 살림
      → query token의 위치가 달라지는 부분도 반영
    - Position-to-position term은 PRE에서 불필요하기 때문에 제거
- `Enhanced Mask Decoder`
  Pre-training 단계에서 Masked Language Model(MLM) 작업을 수행할 때 decoding layer에 절대 위치를 포함하는 Enhanced Mask Decoder를 사용한다. 이는 모델이 [Mask] 토큰을 예측할 때 문장 내 각 단어의 절대적 위치 정보를 효과적으로 활용하여 정확도를 높이는 데 기여했다.
    - Absolute position information의 중요성
- Scale Invariant Fine-Tuning: DeBERTa 모델을 생성하고 SiFT 계층을 모델에 연결
  → 모델의 일반화 능력을 개선하고, 다양한 크기 및 범위의 데이터셋에 대해 모델의 성능을 극대화
    - Adversarial Training은 모델의 일반화에 도움을 준다.
    - Word embedding을 normalize해주고 Perturbation을 추가하자.

5. 이 논문을 토대로 한 Future Work에는 무엇이 있을 수 있는지

- DeBERTa의 disentangled attention 메커니즘의 강점을 활용하여 복잡한 언어적 문맥에서 더 섬세한 이해를 가능하게 하는 연구 진행
  → 다중 언어(multilingual) 데이터셋이나 상호 문화적 텍스트(cross-cultural texts)에 대한 모델링
- 실시간 언어 처리(real-time language processing) 향상
  → 모델의 반응 시간을 개선하여 실시간 대화나 번역 시스템에서의 활용, 속도와 정확성의 균형을 맞추는 데 중점을 둔다.

- 저자원 언어(low-resource languages)에 대한 적용
  → 대량의 데이터가 없는 언어에 대해 DeBERTa를 적용하여 효과적인 언어 모델을 구축
- 인과 관계 학습(cause-and-effect learning)
  DeBERTa 모델을 활용하여 텍스트 내의 원인과 결과 관계를 파악하고 이해하는 능력을 향상
- 감정 분석과 개인화(personalization in sentiment analysis)
  → 개인별 언어 사용 패턴을 분석하여 각 개인의 감정을 더 잘 파악하고 예측할 수 있도록 하는 연구
- 윤리적 NLP 연구(ethical aspects of NLP)
  → DeBERTa를 활용하여 편향이나 왜곡 없이 텍스트를 처리하는 방법론을 개발