**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

**1. 개요 (Abstract)**

* **BERT**(Bidirectional Encoder Representations from Transformers)는 새로운 언어 모델로, **양방향 학습**을 통해 문맥을 더 깊게 이해할 수 있도록 설계됨.
* 기존의 **ELMo**(Peters et al., 2018)나 **GPT**(Radford et al., 2018)와 다르게, 왼쪽과 오른쪽 문맥을 동시에 고려하여 사전 훈련을 수행함.
* 단일 추가 출력 레이어만으로도 다양한 자연어처리(NLP) 작업(예: 질의응답, 문장 추론)에서 **최첨단 성능**을 달성함.
* GLUE 벤치마크에서 **80.5점** (기존 모델 대비 7.7%p 상승), MultiNLI에서 **86.7% 정확도**(4.6%p 증가), SQuAD v1.1에서 **93.2 F1 점수**(1.5%p 증가) 등의 성능을 보여줌.

**2. 서론 (Introduction)**

* **사전 훈련된 언어 모델**은 자연어 처리(NLP) 성능을 향상시키는 데 효과적임.
  + 기존 접근법:
    - **Feature-based** (예: ELMo) → 사전 훈련된 표현을 추가 특성으로 사용
    - **Fine-tuning** (예: GPT) → 모든 가중치를 미세 조정하여 특정 작업에 맞춤
  + 하지만, 기존 언어 모델은 **단방향(unidirectional)** 학습만 가능하여 문맥 이해에 제한이 있었음.
* **BERT는 Masked Language Model(MLM) 기법을 사용하여 양방향 문맥 학습 가능**
  + **MLM**: 입력 문장에서 일부 단어를 가리고, 모델이 이를 예측하도록 훈련
  + **Next Sentence Prediction(NSP)**: 문장 간 관계를 학습

**3. 관련 연구 (Related Work)**

**3.1 비지도 학습 기반 접근법**

* **Word Embedding**: 단어 임베딩(Word2Vec, GloVe)
* **Contextual Embedding**: 문맥을 고려한 임베딩(ELMo, ULMFiT)

**3.2 사전 훈련된 언어 모델**

* **ELMo**: 양방향 LSTM을 사용하지만 독립적으로 훈련 후 병합
* **GPT**: Transformer 기반의 왼쪽에서 오른쪽(Left-to-Right) 언어 모델

**4. BERT 모델 (BERT Model)**

**4.1 모델 구조**

* **Transformer 기반 인코더**(Vaswani et al., 2017) 사용
* 두 가지 버전:
  + **BERT-Base**: 12개 층, 768차원 히든 레이어, 12개 셀프 어텐션 헤드 (총 1.1억 개 파라미터)
  + **BERT-Large**: 24개 층, 1024차원 히든 레이어, 16개 셀프 어텐션 헤드 (총 3.4억 개 파라미터)

**4.2 입력 표현 (Input Representation)**

* **WordPiece Tokenization** (30,000개 토큰 사용)
* 특별한 토큰 사용
  + [CLS] (Classification) → 문장 분류를 위한 토큰
  + [SEP] (Separator) → 문장 구분을 위한 토큰

**5. BERT 사전 훈련 (Pre-training BERT)**

**5.1 Masked Language Model (MLM)**

* 문장에서 **15% 단어를 랜덤하게 마스킹**하여, 이를 예측하도록 훈련
* 기존 **왼쪽-오른쪽** 언어 모델과 다르게, **양방향 문맥 정보 활용 가능**

**5.2 Next Sentence Prediction (NSP)**

* 두 문장이 실제로 연속된 문장인지 예측하도록 학습
* **50%**: 실제로 연결된 문장(IsNext)
* **50%**: 랜덤한 문장(NotNext)
* 질의응답(QA) 및 자연어 추론(NLI) 성능 향상에 기여

**6. BERT 미세 조정 (Fine-tuning BERT)**

* 사전 훈련된 BERT 모델을 사용해 **하위 작업에 맞춰 전체 가중치를 조정**
* 다양한 자연어처리(NLP) 작업 적용 가능
  + **문장 분류 (GLUE, SWAG)**
  + **질의응답 (SQuAD)**
  + **이름 인식(NER, CoNLL-2003)**

**7. 실험 및 결과 (Experiments & Results)**

**7.1 GLUE 벤치마크 결과**

* BERT는 기존 모델 대비 **최고 성능을 달성**
* **BERT-Large**가 **OpenAI GPT**보다 모든 GLUE 태스크에서 우수한 성능 기록

| **Model** | **MNLI** | **QNLI** | **SST-2** | **CoLA** | **MRPC** | **RTE** | **STS-B** | **평균** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| GPT | 82.1 | 87.4 | 91.3 | 45.4 | 82.3 | 56.0 | 80.0 | 75.1 |
| **BERT-Base** | 84.6 | 90.5 | 93.5 | 52.1 | 88.9 | 66.4 | 85.8 | 79.6 |
| **BERT-Large** | **86.7** | **92.7** | **94.9** | **60.5** | **89.3** | **70.1** | **86.5** | **82.1** |

**7.2 SQuAD (질의응답) 결과**

* **BERT-Large(93.2 F1)** → 기존 모델보다 **+1.5 F1 증가**
* **SQuAD 2.0에서도 최고 성능 달성 (83.1 F1, +5.1 F1 증가)**

**8. 분석 (Ablation Study)**

**8.1 사전 훈련 태스크 중요성**

* **NSP 제거 시 성능 하락** (특히 질의응답 및 자연어추론)
* **MLM이 단방향 모델보다 성능 우수**

**8.2 모델 크기 영향**

* **더 깊고 큰 모델이 성능 향상에 기여**
* 특히 **작은 데이터셋에서도 큰 모델이 효과적**

**9. 결론 (Conclusion)**

* **BERT는 기존 모델보다 더 강력한 언어 표현 모델**
* **양방향 문맥 정보 활용 가능 → 다양한 NLP 작업에서 최고 성능**
* **사전 훈련된 가중치만으로도 고품질 NLP 모델 구축 가능**
* **공개된 코드 및 사전 훈련 모델 활용 가능** ([GitHub](https://github.com/google-research/bert))

**🔑 BERT 논문의 핵심 포인트 🔑**

✔ **양방향 학습(Bidirectional Learning) 도입** → 기존 단방향 언어 모델의 한계를 극복하여 문맥 이해력 대폭 향상!

✔ **Masked Language Model (MLM) 적용** → 왼쪽 & 오른쪽 문맥을 모두 활용하여 학습, 더 정교한 언어 표현 가능!

✔ **Next Sentence Prediction (NSP) 기법 추가** → 문장 간 관계를 학습하여 질의응답(QA), 자연어 추론(NLI) 성능 향상!

✔ **Fine-tuning 방식의 강력함** → 사전 훈련된 모델을 간단한 미세 조정만으로 다양한 NLP 작업에 활용 가능!

✔ **다양한 NLP 벤치마크에서 최고 성능 기록** → GLUE, SQuAD 등 여러 과제에서 기존 모델 대비 **압도적 성능 개선!**

✔ **공개된 코드 & 사전 훈련 모델 활용 가능** → 연구자 & 개발자가 쉽게 활용하여 NLP 기술 발전에 기여! ([GitHub 링크](https://github.com/google-research/bert))