**Improving Language Understanding by Generative Pre-Training**

**1. 개요 (Abstract)**

* 자연어 이해(NLU) 과제는 질의응답(QA), 의미적 유사성 판단, 문서 분류 등 다양한 분야에서 활용됨.
* 대부분의 NLP 모델은 \*\*지도 학습(Supervised Learning)\*\*이 필요하지만, **라벨이 부족**하여 학습이 어렵다는 문제가 있음.
* 이를 해결하기 위해 \*\*비지도 사전 훈련(Generative Pre-Training, GPT)\*\*을 도입하고, 이후 각 태스크에 맞춰 **미세 조정(Fine-Tuning)** 수행.
* 기존 모델과 달리, **태스크별 아키텍처 변경 없이** 효과적인 전이 학습을 수행할 수 있도록 설계됨.
* **GLUE 벤치마크 기준, 12개 태스크 중 9개에서 기존 SOTA(SOTA: 최고 성능) 모델을 능가**함.
* 특히 \*\*상식적 추론(Stories Cloze Test, +8.9%), 질의응답(RACE, +5.7%), 텍스트 포함관계(MultiNLI, +1.5%)\*\*에서 성능 개선을 보임.

**2. 서론 (Introduction)**

* **자연어 이해 모델의 발전에는 대량의 라벨링 데이터가 필요하지만, 데이터 부족이 큰 장애물**
* 사전 훈련(Pre-training)은 **비지도 학습을 통해 언어적 특징을 학습**하고, 이후 **지도 학습으로 특정 태스크에 적응**하는 방법
* 기존 연구들(ELMo, ULMFiT 등)과 비교하여 \*\*GPT는 Transformer 기반의 단방향 언어 모델(Unidirectional LM)\*\*을 활용하여 사전 훈련 수행
* 사전 훈련된 모델을 **각 태스크에 미세 조정**하는 방식으로, 별도의 구조 변경 없이 전이 학습 가능

**3. 관련 연구 (Related Work)**

**3.1 NLP에서 반지도 학습(Semi-Supervised Learning)**

* 이전에는 **Word2Vec, GloVe** 등의 단어 수준 임베딩을 활용했지만, 문장 수준 정보는 충분히 반영되지 않았음.
* ELMo, ULMFiT 같은 모델이 등장하며 **문맥을 고려한 동적 단어 표현**이 가능해짐.
* 하지만, LSTM 기반이라 **장기 의존성(Long-range Dependency) 처리 한계**가 존재했음.

**3.2 비지도 사전 훈련(Unsupervised Pre-training)**

* Dai & Le (2015), Howard & Ruder (2018) 등의 연구에서는 **LSTM 기반 언어 모델을 사전 훈련 후 미세 조정**하는 방법을 제안.
* 하지만, **LSTM은 단기 의존성에 강하고 장기 문맥 이해에는 한계가 있음**.
* GPT는 Transformer 기반으로, **장기 문맥 이해 및 전이 학습 성능을 개선**.

**4. GPT 모델 구조 (Framework: Generative Pre-Training GPT)**

**4.1 사전 훈련(Unsupervised Pre-training) 단계**

* **목표:** 대량의 비지도 텍스트 데이터로 **일반적인 언어 표현 학습**
* **방법:**
  + Transformer 기반 **단방향 언어 모델(Left-to-Right Language Model)** 사용
  + 입력 단어 시퀀스에서 **다음 단어 예측(Next-Word Prediction) 학습**
  + 학습 손실 함수: L1(U)=∑ilog⁡P(ui∣ui−k,...,ui−1;Θ)L\_1(U) = \sum\_{i} \log P(u\_i | u\_{i-k}, ..., u\_{i-1}; \Theta)L1​(U)=i∑​logP(ui​∣ui−k​,...,ui−1​;Θ)
  + 여기서 kkk는 컨텍스트 크기, Θ\ThetaΘ는 모델의 파라미터
* **데이터셋:** BooksCorpus (7,000권 이상의 책, 긴 문맥 유지 가능)

**4.2 미세 조정(Supervised Fine-tuning) 단계**

* **목표:** 사전 훈련된 모델을 특정 태스크에 맞게 조정
* **방법:**
  + 사전 훈련된 모델을 활용하여 각 태스크별로 supervised 학습 진행
  + 출력층만 추가하여 최소한의 구조 변경으로 적용
  + 손실 함수: L2(C)=∑(x,y)log⁡P(y∣x1,...,xm)L\_2(C) = \sum\_{(x,y)} \log P(y | x\_1, ..., x\_m)L2​(C)=(x,y)∑​logP(y∣x1​,...,xm​)
  + **추가적인 언어 모델링 손실을 포함**하여 학습 성능 향상 L3(C)=L2(C)+λ⋅L1(C)L\_3(C) = L\_2(C) + \lambda \cdot L\_1(C)L3​(C)=L2​(C)+λ⋅L1​(C)

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

**5.1 평가 데이터셋**

* **GLUE 벤치마크** 포함 12개 NLP 태스크 실험
  + 자연어 추론(NLI) → **SNLI, MultiNLI, SciTail**
  + 질의응답(QA) → **RACE, Story Cloze Test**
  + 문장 유사도 → **MSR Paraphrase, STS-B**
  + 텍스트 분류 → **SST-2, CoLA**

**5.2 주요 성능 개선**

| **Task** | **기존 SOTA** | **GPT (ours)** | **성능 향상** |
| --- | --- | --- | --- |
| **상식적 추론 (Story Cloze Test)** | 77.6 | **86.5** | **+8.9%** |
| **질의응답 (RACE)** | 53.3 | **59.0** | **+5.7%** |
| **자연어 추론 (MultiNLI)** | 80.6 | **82.1** | **+1.5%** |
| **문장 유사도 (STS-B)** | 81.0 | **82.0** | **+1.0%** |
| **문법 판단 (CoLA)** | 35.0 | **45.4** | **+10.4%** |
| **GLUE 종합 점수** | 68.9 | **72.8** | **+3.9%** |

**5.3 Ablation Study (성능 분석)**

* Transformer 구조가 LSTM보다 **긴 문맥 이해 능력이 뛰어남**
* **사전 훈련을 생략할 경우 성능 14.8% 감소**
* **보조 언어 모델링 손실 추가 시 성능 증가**

**6. 결론 (Conclusion)**

✔ **비지도 사전 훈련 + 지도 미세 조정의 강력한 조합** → 태스크별 데이터 부족 문제 해결 가능!

✔ **Transformer 기반 단방향 언어 모델 사용** → LSTM보다 **장기 문맥 이해 능력**이 뛰어남!

✔ **다양한 NLP 태스크에서 최고 성능 기록** → GLUE 9개 태스크에서 SOTA 모델 능가!

✔ **태스크별 모델 구조 변경 없이 적용 가능** → 간단한 미세 조정만으로 전이 학습 수행!

**🔑 GPT 논문의 핵심 포인트 🔑 -> 이 논문이 GPT 시리즈의 시작점이라 매우 중요함!!!**

✔ **비지도 사전 훈련 + 지도 미세 조정(Generative Pre-Training + Fine-Tuning)**  
→ 대량의 라벨 없는 텍스트에서 언어 패턴을 학습한 후, 각 태스크에 맞게 최소한의 조정만으로 적용 가능!

✔ **Transformer 기반 단방향 언어 모델(Left-to-Right LM) 사용**  
→ 기존 LSTM보다 **더 긴 문맥을 학습 가능**하여 전이 학습 성능 대폭 향상!

✔ **태스크별 구조 변경 없이 다양한 NLP 작업 적용 가능**  
→ 별도의 아키텍처 수정 없이 질의응답(QA), 자연어 추론(NLI), 문장 유사도, 문서 분류 등에서 **최고 성능** 달성!

✔ **GLUE 벤치마크 9개 태스크에서 기존 최고 성능(SOTA) 모델 능가**  
→ 특히 \*\*상식적 추론(📈 +8.9%), 질의응답(📈 +5.7%), 문법 판단(📈 +10.4%)\*\*에서 큰 개선!

✔ **LSTM 대비 Transformer의 장기 문맥 이해력 증명**  
→ Ablation Study 결과, Transformer는 **LSTM보다 전이 학습에 효과적**이며, **사전 훈련이 없으면 성능 14.8% 감소!**

✔ **자연어 처리(NLP) 패러다임 전환의 시작점**  
→ 이 논문을 기반으로 이후 BERT, GPT-2, GPT-3, ChatGPT 같은 모델들이 등장하며 **자연어 이해 기술의 새로운 장을 열게 됨!**