GPT-1: Improving Language Understanding by Generative Pre-Training

기간	@04/30/2025 → 05/06/2025
주차	9주차
논문	GPT-1.pdf
상태	완료
예습/복습	예습과제
참고 자료	참고 블로그1 참고 블로그2

0. Abstract

- 배경 : 자연어 이해 → 텍스트의 함의 판단, 질문 응답, 의미 유사도 평가, 문서 분류 등 매우 다양한 과제 포함.
    - 대규모 unlabeled text corpora 풍부하게 존재하지만, 이러한 특정 과제를 학습하기 위한 labeled 데이터는 부족 ⇒ 판별 기반(discriminatively trained) 모델들이 충분한 성능 내기 어려움.
- **본 논문은 대규모 unlabeled text corpus에 대해 언어 모델을 생성 기반으로 사전학습(pre-training)하고, 이후 각 과제별로 판별 방식(fine-tuning)으로 미세 조정하는 방식**이 위 과제에서 큰 성능 향상 가져올 수 있음을 보여줌.
- 기존 접근법과 달리, 미세 조정 시 **task-aware input transformation 사용**함으로써 모델 구조 변경 최소화하면서도 효과적인 전이를 달성함.
- 이 접근 방식이 다양한 자연어 이해 벤치마크에서 효과적임을 입증함.
    - 각 과제마다 특별히 설계된 아키텍처 사용하는 기존 판별 학습 모델보다 뛰어난 성능 보임. (연구된 12개 과제 중 9개에서 최신 성능 크게 상회)
    - Stories Cloze Test에서 8.9%, 질문 응답(RACE)에서 5.7%, 텍스트 함의(MultiNLI)에서 1.5%의 절대 성능 향상 달성.

1. Introduction

논문이 다루는 분야

- **비정형(raw) 텍스트로부터 효과적으로 학습**하는 능력 → 자연어 처리(NLP)에서 supervised 학습에 대한 의존을 줄이는 데 있어 매우 중요.

해당 task에서 기존 연구 한계점

- 대부분 딥러닝 기반 방법들 → 대규모 수작업 라벨 데이터를 필요로 하는데, 이는 라벨링 자원이 부족한 많은 도메인에서 모델 적용 어렵게 함.

    - 이러한 경우, unlabeled 데이터를 통해 언어 정보를 활용할 수 있는 모델 <= 새로운 라벨 수집하는 데 소요되는 시간과 비용 줄이는 유용한 대안될 수 있음.
- supervised 정보가 풍부한 경우에도 비지도 방식으로 좋은 표현을 학습하는 것 → 성능 향상시킬 수 있음.
    - 가장 설득력 있는 증거 → 다양한 NLP 과제에서 성능 향상에 기여한 **pre-trained word embeddings**의 광범위한 활용.
- But, 단어 수준을 넘는 표현을 unlabeled 데이터로부터 학습하는 것은 2가지 주요 이유로 어려움.
    1. 어떤 종류의 최적화 목적 함수가 전이에 유용한 텍스트 표현을 학습하는 데 가장 효과적인지 명확하지 않음.
        - 최근 연구들은 언어 모델링, 기계 번역, 담화(coherence) 등 다양한 목적 함수 실험했으며, 각 방법이 특정 과제에서는 다른 방법보다 더 나은 성능 보임.
    2. 학습된 표현을 목표 과제에 어떻게 전이시킬 것인지에 대해서도 합의된 방식이 없음.
        - 기존 기법들은 모델 구조를 과제에 맞게 바꾸거나, 복잡한 학습 shemes(계획) 또는 보조 학습 목적 함수를 추가하는 방식 등 사용함.

    ⇒ 이러한 불확실성은 효과적인 semi-지도 학습 접근 방식을 개발하는 데 걸림돌이 됨.

논문의 contributions

- 본 논문에서는 **unsupervised pre-training과 supervised fine-tuning을 결합한 반지도 학습(semi-supervised learning) 방식** 사용하여 언어 이해 과제 다룸.
- 목표 : 다양한 과제에 **최소한의 adaptation으로 전이 가능한 범용 표현(universal representation) 학습**하는 것!
    - 대규모 unlabeled text corpus와 여러 개의 labeled 데이터셋(타깃 과제) 사용할 수 있다고 가정. (이때, unlabeled text corpus와 타깃 과제가 동일한 도메인일 필요는 없음.)
    - **두 단계의 학습 절차**
        1. **사전 학습:** unlabeled 데이터 기반으로 언어 모델링 목적 함수 사용하여 신경망 초기 파라미터 학습함.
        2. **미세 조정:** 학습된 파라미터를 지도학습 목적 함수 이용하여 타깃 과제에 맞게 조정함.

- 모델 아키텍처 : **Transformer**
  - 기계 번역, 문서 생성, 구문 분석 등 다양한 작업에서 강력한 성능 보여줌.
  - 순환 신경망(RNN)보다 장기적인 의존 관계를 더 효과적으로 처리할 수 있는 구조화된 memory 제공하여 다양한 작업에 대해 강건한 전이 성능 보임.
- 전이 학습 → **traversal-style 방식** 차용
  - 입력을 하나의 연속된 토큰 시퀀스로 변환함으로써 구조화된 입력 처리함.

    ⇒ 기존 모델 구조 거의 변경하지 않고도 효과적인 미세 조정 가능함.
- 자연어 추론, 질문 응답, 의미 유사도, 텍스트 분류하는 4가지 언어 이해 과제에서 해당 접근법 평가함.
  - 각 과제를 위해 특별히 설계된 구조 사용하는 기존 모델보다 더 나은 성능 보임.
    - 상식 추론 (Stories Cloze Test): +8.9%
    - 질문 응답 (RACE): +5.7%
    - 텍스트 함의 (MultiNLI): +1.5%
    - GLUE 벤치마크 평균: +5.5%
- 사전학습된 모델의 **zero-shot 성능 분석**하여, 모델이 다양한 downstream 과제에 유용한 언어 지식 획득함을 보임.

2. Related Work

Semi-supervised learning for NLP

- 반지도 학습(Semi-supervised learning)
  - 시퀀스 라벨링, 텍스트 분류 등의 과제에 널리 활용됨.
  - 초기 접근법) unlabeled 데이터 활용하여 단어 혹은 구 수준의 통계 계산한 후, 이를 지도 모델의 feature로 사용하는 방식
  - 이후 연구들) unlabeled corpus에서 훈련된 word embedding의 장점 보여주었고, 이는 다양한 과제에서 성능 향상시킴.
  - 한계) 주로 단어 수준의 정보 전이에 국한됨.
  - 최근에는 단어를 넘어서 **구, 문장 수준의 임베딩**을 unlabeled 데이터에서 학습하려는 시도 이어짐. → 이러한 임베딩은 다양한 과제에서 문장을 벡터 표현으로 변환하여 사용됨.

Unsupervised pre-training

- 비지도 사전학습(Unsupervised Pre-training)
  - 반지도 학습의 한 형태, 목표 → **좋은 초기화 지점(good initialization)** 찾는 것!
  - 초기 연구들) 이미지 분류 또는 회귀 문제에서 적용
  - 이후 연구들) 사전 학습이 **정규화(regularization) 효과**를 가져와 일반화 성능을 높이는 데 기여함을 보임.
  - 최근) 이미지 분류, 음성 인식, entitiy disambiguation, 기계 번역의 분야에서 비지도 사전 학습 활용되고 있음.
- 본 연구와 가장 가까운 선행 연구 → **언어 모델링 목적 함수로 신경망 사전 학습하고, 이를 타깃 과제에 지도 학습으로 미세 조정**하는 방식
  - Dai et al. & Howard & Ruder) 이 방식을 텍스트 분류에 사용했지만, LSTM 기반 모델 사용했기에 단기 문맥만 처리 가능하다는 제약.
  - 반면, 본 논문은 **Transformer 사용하여 더 긴 문맥 구조 처리 가능**.
    - 자연어 추론, paraphrase 검출, 이야기 완성 등 더 다양한 과제에 효과적임을 실험으로 입증함.

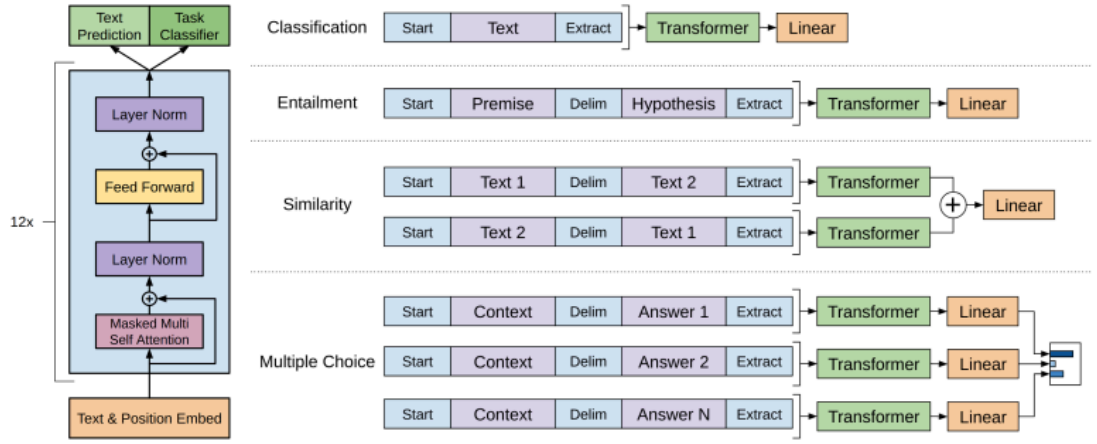
- 기존 접근법 → 각 과제에 대해 많은 파라미터 새로 추가해야 함.
- 본 논문 → 아키텍처 변경 없이 전이 가능!

Auxiliary training objectives

- 보조 학습 목적 함수(Auxiliary Training Objectives)
  - Collobert & Weston) 품사 태깅, 개체명 인식, 언어 모델링 등 다양한 NLP 작업을 보조 목표로 추가하여 성능 향상시킴.
  - 최근 연구) 주된 목적 함수에 언어 모델링 추가하여 시퀀스 라벨링 성능 향상시킴.

- 실험에서 보조 목적 함수 사용했지만, **비지도 사전학습만으로도 이미 많은 언어적 특성 학습함**을 확인함.

3. 제안 방법론



3.1 Unsupervised pre-training

- 대규모 text corpus에서 언어 모델 학습
- 주어진 비지도 토큰 시퀀스 $U = \{u_1, u_2, \ldots, u_n\}$에 대해, 언어 모델링 목표 함수 사용하여 다음의 likelihood 최대화함.

$$L_1(U) = \sum_i logP(u_i|u_{i-k}, ..., u_{i-1}; \Theta)$$

  - k : 문맥 윈도우 크기
  - P : 다음 토큰의 조건부 확률
  - Θ : 모델 파라미터(stochastic gradient descent로 학습됨)
- 모델 구조
  - 다층 **Transformer 디코더(Decoder-only) 사용**
  - 입력 토큰에 대해 멀티헤드 self-attention 적용
  - 각 위치마다 개별적인 feed-forward 레이어 거쳐 출력 분포 계산함.

$$h_0 = UW_e + W_p$$
$$h_t = transfomer\_block(h_{t-1}) \forall i \in [1, n]$$
$$P(u) = softmax(h_n W_e^T)$$

  - W_e : 토큰 임베딩 행렬
  - W_p : 위치 임베딩 행렬
  - n : Transformer 레이어 수

3.2 Supervised fine-tuning

- 사전 학습 마친 후, 해당 모델을 **라벨이 있는 과제**에 맞춰 미세 조정
- 주어진 데이터셋 $C = \{(x, y)\}$에 대해 입력 시퀀스 $x_1, \ldots, x_m$과 라벨 y 사용함.
- 입력 시퀀스를 모델에 넣고 마지막 Transformer 블록의 출력을 선형 계층에 통과시켜 예측 확률 계산

$$P(y|x_1, ..., x_m) = softmax(h_m^l W_y)$$

- 이에 따라 미세조정 단계의 손실 함수는 다음과 같음.

$$L_2(C) = \sum_{(x,y)} logP(y|x_1, ..., x_m)$$

- 추가로, fine-tuning 단계에서 언어 모델링 손실 L_1을 보조 목적 함수로 활용하면 다음 2가지 이점이 있었음.
  1. 일반화 성능 향상
  2. 학습 속도 증가
- 따라서, 최종 fine-tuning 손실 함수 다음과 같음.

$$L_3(C) = L_2(C) + \lambda \cdot L_1(C)$$

  - λ : 보조 손실의 가중치
  - fine-tuning 시 새로 추가되는 파라미터는 출력층 W_y와 특별 토큰(시작, 종료)의 임베딩뿐.

3.3 Task-specific input transformations

- 사전학습된 모델은 연속된 텍스트 시퀀스 입력에 최적화되어 있으므로 일부 과제에서는 입력 구조 변환해야 함.

  ⇒ **traveral-style 접근법** 사용하여 구조화된 입력을 하나의 시퀀스로 변환!
- **변환 예시**
  1. **Textual Entailment(텍스트 함의 판단)**
     - 전제(p)와 가설(h)을 구분 토큰($)으로 이어 붙여 입력
  2. **Semantic Similarity(문장 유사도 평가)**

- 문장 순서에 의미가 없기 때문에 두 가지 순서를 각각 처리한 후 결과를 더하여 출력층에 전달

$$h_{ml}^{(1)} + h_{ml}^{(2)} \rightarrow softmax$$

3. **Question Answering & Commonsense Reasoning**
   - 문맥(z), 질문(q), 선택지 답변 a_k를 다음처럼 결합
     - [z; q; $; a_k]
   - 각 조합에 대해 softmax 적용하여 가장 높은 확률 갖는 답 선택

4. 실험 및 결과

4.1 Setup

- **사전 학습 데이터**
  - BooksCorpus 사용 (약 7,000권의 책, 장르 다양: 판타지, 로맨스 등)
  - 문장 순서를 유지한 긴 텍스트 시퀀스 포함 → 언어 모델링에 적합
  - alternative 데이터셋인 1B Word Benchmark(ELMo와 유사)는 문장 단위로 셔플되어 장기 문맥 학습에 불리
- **모델 구조**
  - 12층 Transformer decoder
  - 768차원 hidden state, 12개의 self-attention 헤드
  - Position-wise FFN : 3072차원
  - **옵티마이저** : Adam (최대 learning rate 2.5e-4, cosine decay)
  - **학습 epoch** : 100, 배치당 512 토큰씩 64개 시퀀스
  - **Dropout** : 0.1 (embedding, attention, residual 등에 적용)
  - GELU 활성화 함수
  - Byte Pair Encoding(BPE) vocabulary, 40,000개 토큰
  - **텍스트 전처리** : ftfy, spaCy 토크나이저
- **Fine-tuning 설정**
  - 대부분 동일한 하이퍼파라미터 유지

- **learning rate** : 6.25e-5
- **batch size**: 32
- 3 epoch만으로 충분히 수렴
- 학습 초반 0.2% warm-up 사용
- 보조 손실 가중치 $\lambda = 0.5$

4.2 Supervised fine-tuning

Task	Datasets
Natural language inference	SNLI [5], MultiNLI [66], Question NLI [64], RTE [4], SciTail [25]
Question Answering	RACE [30], Story Cloze [40]
Sentence similarity	MSR Paraphrase Corpus [14], Quora Question Pairs [9], STS Benchmark [6]
Classification	Stanford Sentiment Treebank-2 [54], CoLA [65]

- 4가지 유형의 자연어 이해 과제에서 GPT 모델 fine-tuning 함.

A.
Natural Language Inference (자연어 추론)

- 데이터셋
  - **MultiNLI** : 다중 장르 포함
  - **QNLI** : 질문과 문장을 매칭
  - **RTE** : Recognizing Textual Entailment
  - **SciTail** : 과학적 문장 기반 추론
- 결과

Method	MNLI-m	MNLI-mm	SNLI	SciTail	QNLI	RTE
ESIM + ELMo [44] (5x)	-	-	89.3	-	-	-
CAFE [58] (5x)	80.2	79.0	89.3	-	-	-
Stochastic Answer Network [35] (3x)	80.6	80.1	-	-	-	-
CAFE [58]	78.7	77.9	88.5	83.3		
GenSen [64]	71.4	71.3	-	-	82.3	59.2
Multi-task BiLSTM + Attn [64]	72.2	72.1	-	-	82.1	**61.7**
Finetuned Transformer LM (ours)	**82.1**	**81.4**	**89.9**	**88.3**	**88.1**	56.0

  - 기존 구조보다 GPT 모델이 **더 긴 문맥을 잘 처리**하여 전반적으로 좋은 성능을 보임
  - 특히 MultiNLI에서 기존 최고 성능보다 **1.5% 향상**

B.

Question Answering and commonsense reasoning (질문 응답 및 상식 추론)

- 데이터셋
  - **RACE** : 고난이도 영어 독해 질문 (중·고등학생용 시험 문제)
  - **ROCStories Cloze** : 상식 기반 이야기 완성 문제
- 결과

Method	Story Cloze	RACE-m	RACE-h	RACE
val-LS-skip [55]	76.5	-	-	-
Hidden Coherence Model [7]	77.6	-	-	-
Dynamic Fusion Net [67] (9x)	-	55.6	49.4	51.2
BiAttention MRU [59] (9x)	-	60.2	50.3	53.3
Finetuned Transformer LM (ours)	**86.5**	**62.9**	**57.4**	**59.0**

  - RACE에서는 기존 최고 성능 대비 **5.7% 절대 향상**
  - ROC Cloze에서는 **8.9% 향상**, 특히 단어 수준 선택지보다 문장 수준 선택지에 효과적

C. **Semantic Similarity (문장 유사도 평가)**

- 데이터셋
  - **MRPC**: 패러프레이즈 검출 (Microsoft Research)
  - **QQP**: Quora 유사 질문 판단
  - **STS-B**: Semantic Textual Similarity Benchmark (연속값 회귀 문제)
- 결과
  - GPT 모델은 입력 순서에 대한 민감성을 보였기 때문에 **양방향 입력을 각각 평가한 후 평균을 사용**
  - 회귀 문제(STS-B)는 분류기 대신 평균 스코어 예측 구조 사용
  - STS-B에서 기존 최고 모델 대비 **1.0 포인트 향상**

D.

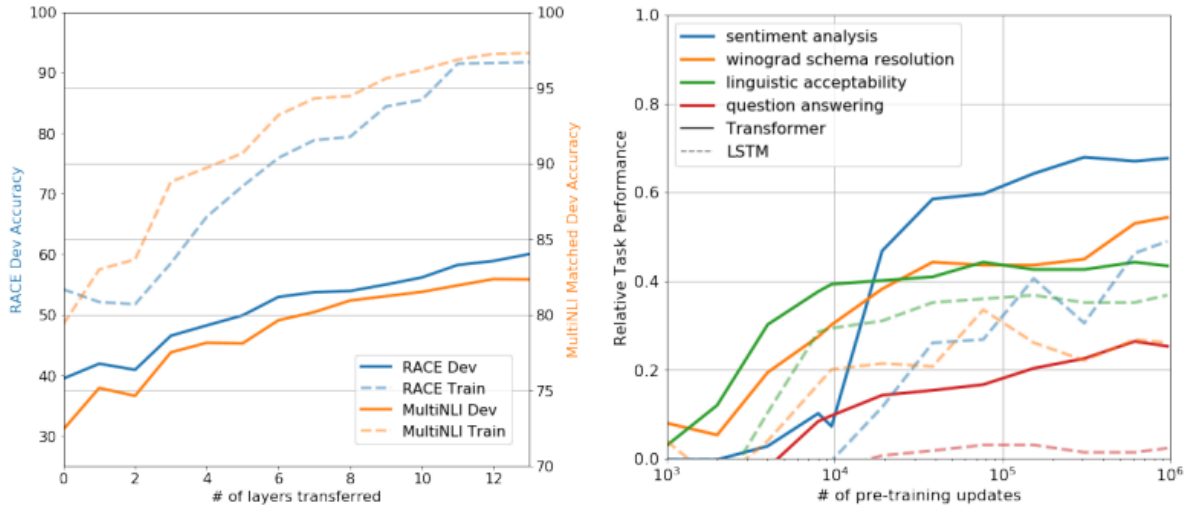
Classification (문장 분류)

- 데이터셋
  - **SST-2**: 감성 분석 (Stanford Sentiment Treebank)
  - **CoLA**: 문법성 판단 (Corpus of Linguistic Acceptability)
- 결과
  - SST-2에서는 기존 최고 모델 수준 성능
  - CoLA에서는 **10% 이상 향상된 Matthews correlation 점수 (45.4)** → 문법 판단에 매우 효과적

Method	Classification		Semantic Similarity			GLUE
	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	
Sparse byte mLSTM [16]	-	**93.2**	-	-	-	-
TF-KLD [23]	-	-	**86.0**	-	-	-
ECNU (mixed ensemble) [60]	-	-	-	81.0	-	-
Single-task BiLSTM + ELMo + Attn [64]	35.0	90.2	80.2	55.5	66.1	64.8
Multi-task BiLSTM + ELMo + Attn [64]	18.9	91.6	83.5	72.8	63.3	68.9
Finetuned Transformer LM (ours)	**45.4**	91.3	82.3	**82.0**	**70.3**	**72.8**

5. 분석(Analysis)

Impact of number of layers transferred



- unsupervised pre-training 모델을 supervised target task으로 전이할 때, transfer layer의 개수가 미치는 영향 측정
  - 사용한 transfer layer의 개수가 많을수록 성능이 좋아짐.
  - pre-trained 모델의 정보를 많이 사용할수록 fine-tuning 모델도 성능이 좋아짐.
  - pre-trained 모델의 각 layer가 target task 해결하는데 유용한 기능 가지고 있음.

Zero-shot Behaviors

- 왜 pre-training이 모델 성능 향상에 도움을 주는가?
  - pre-training을 많이 진행할수록 fine-tuning 이후의 성능이 좋아짐.
  - underlying generative model이 pre-training을 통해 다양한 task를 수행하는데 필요한 **많은 부분을 이미 학습한다**는 것을 의미.
  - 또한 Transformer 구조가 LSTM보다 underlying generative model의 구조로 더 적합함.

Ablation Studies

Method	Avg. Score	CoLA (mc)	SST2 (acc)	MRPC (F1)	STSB (pc)	QQP (F1)	MNLI (acc)	QNLI (acc)	RTE (acc)
Transformer w/ aux LM (full)	74.7	45.4	91.3	82.3	82.0	**70.3**	**81.8**	**88.1**	**56.0**
Transformer w/o pre-training	59.9	18.9	84.0	79.4	30.9	65.5	75.7	71.2	53.8
Transformer w/o aux LM	**75.0**	**47.9**	**92.0**	**84.9**	**83.2**	69.8	81.1	86.9	54.4
LSTM w/ aux LM	69.1	30.3	90.5	83.2	71.8	68.1	73.7	81.1	54.6

1. fine-tuning에서 auxiliary LM objective을 포함하지 않는 모델의 성능
   - **auxiliary objective는 NLI과 QQP task에서 유용함**
   - **dataset이 클수록 auxiliary objective의 이득을 더 받음**
2. Transformer **vs** single layer 2048 unit LSTM
   - Transformer 대신 **LSTM을 사용하면 평균 score가 5.6% 하락**
   - **MRPC에 대해서만 LSTM이 더 좋은 성능**을 보임
3. pre-training하지 않은 모델과 비교
   - 모든 task에 있어서 **pre-training을 하지 않고** 바로 모델링한 경우 **성능이 저하**
   - 평균 score가 14.8% 하락

6. 결론(Conclusion)

- 본 논문에서는 **범용적인 자연어 이해 모델을 구축하기 위한 새로운 접근법**을 제안
- **주요 특징**
  1. **대규모 unlabeled 텍스트로 사전학습**
  2. **구조 변경 없이 과제별 fine-tuning**
  3. **Transformer 아키텍처**를 통한 긴 문맥 학습
  4. 다양한 과제에서 일관된 성능 향상
- **성과**
  - 12개 과제 중 9개에서 **최신 최고 성능 달성**
  - GLUE 벤치마크 기준, 기존 최고 성능 대비 **5.5 포인트 향상**
  - 적은 양의 라벨 데이터로도 강력한 성능 발휘 → **semi-supervised learning의 가능성 실증**
- **발전 가능성**
  - 더 큰 사전학습 데이터셋 사용
  - 다양한 언어·도메인으로 확장
  - 복잡한 reasoning이나 multi-hop QA 과제에 적용