GPT-2 구축하기:
자연어처리 태스크를 위한 Fine-tuning 전략 비교

감정 분석 · 패러프레이즈 탐지 · 시 생성 태스크 기반 실험

[CSC4027] 자연어처리개론 | 김광일 교수님

안하현 | 인공지능전공 2022113162
이지현 | AI융합학부 2022113134
전희연 | 컴퓨터공학과 2021113415

dongguk UNIVERSITY

Contents

연구 배경

최근 자연어처리(NLP) 분야에서는
사전학습 언어모델(PLM)의 발전으로
감정 분석, 문장 분류, 생성 과제 등
다양한 태스크에서 성능이 향상되고 있다.
특히 **GPT-2는 decoder-only 구조의**
autoregressive 모델로 문맥 기반의
자연스러운 텍스트 생성을 수행하며
텍스트 생성 중심 태스크에 폭넓게 활용된다.

프로젝트 목표

본 프로젝트에서는 GPT-2 모델을 구현하고
감정 분석, 패러프레이즈 탐지, 시 생성 등
다양한 NLP 태스크에 적용하였다.
또한, **fine-tuning 전략을 비교 실험**하여
GPT-2의 구조적 특성과 태스크별 적용
가능성을 실증적으로 고찰하고자 하였다.

PART I – GPT-2 구현하기

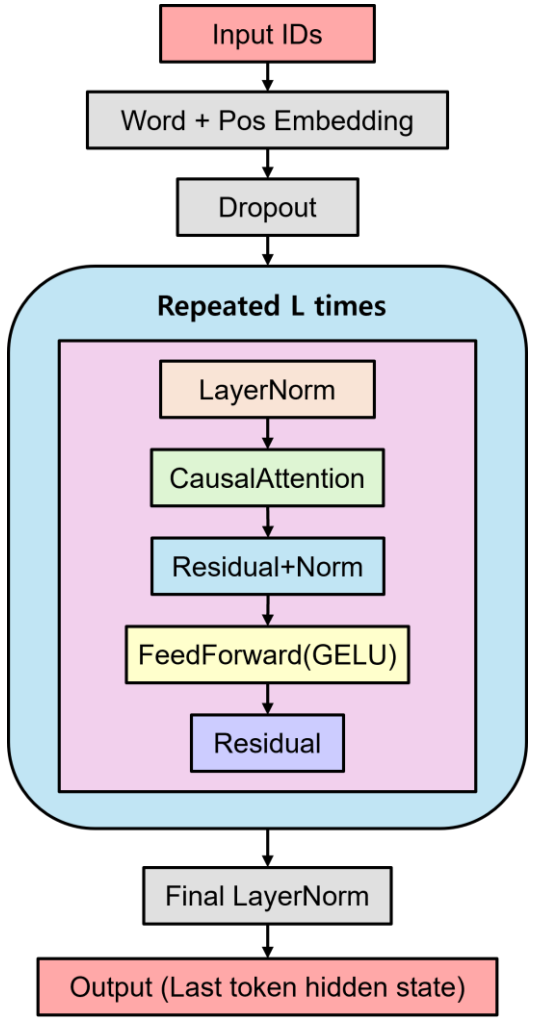
GPT-2 모델 구조 개요

본 프로젝트는 **OpenAI의 GPT-2 구조**를 기반으로,
Transformer의 핵심 원리를 직접 구현하고
자연어처리 태스크에 활용할 수 있도록 재설계하였다.

GPT-2는 **decoder-only 구조의 언어모델**로,
문맥을 고려한 텍스트 생성을 수행하며
분류와 생성 태스크 모두에 적용 가능하다.

구현한 GPT-2는 다음과 같은 순서로 입력을 처리한다:

Input IDs → Embedding → Dropout → Transformer Block × L
→ Final LayerNorm → Output



Transformer 레이어 구성

① LayerNorm

→ 입력값을 정규화하여 학습 안정성을 높임

② Masked Multi-head Self-Attention

→ 문맥 정보를 여러 관점에서 처리하며

미래 토큰을 가리는 Causal Mask 적용

③ Residual + LayerNorm

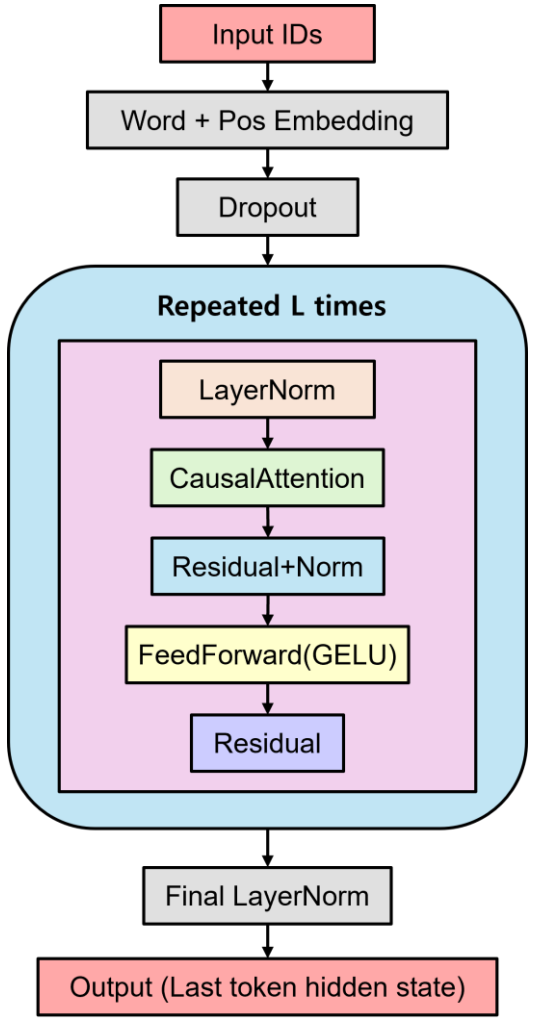
→ 정보 흐름을 보존하며 학습 시 gradient 소실 방지

④ FeedForward (2-layer MLP + GELU)

→ 각 위치의 표현을 비선형적으로 확장하여 모델 표현력 향상

⑤ Residual 연결

→ 이전 정보와 현재 출력을 더해 정보 손실 방지



PART I – 감정 분석 실험

데이터셋 소개

Stanford Sentiment Treebank (SST-5)

- 영화 리뷰에 기반한 5단계 감정 레이블
  - 0: Very Negative
  - 1: Negative
  - 2: Neutral
  - 3: Positive
  - 4: Very Positive
- 총 문장 수: 약 8,500개
- 모델은 "문장 + 감정 예측" 형태로 분류 학습

Korean IMDB 리뷰 (CFIMDB)

- 한국어 영화 리뷰에 기반한 2단계 감정 레이블
  - 0: Negative
  - 1: Positive
- 총 문장 수: 약 12,000개
- 실제 한국어 리뷰에 GPT-2를 적용하는 데 사용됨

BASELINE 학습

학습 방식 및 하이퍼파라미터

- GPT-2의 마지막 토큰 hidden state를 분류기로 연결
- 전체 파라미터를 학습하는 Full Fine-Tuning 방식 사용
- CrossEntropyLoss 적용
- Optimizer: AdamW

항목	SST-5	CFIMDB
Batch Size	64	8
Epochs	6	6
Learning Rate	2e-5	2e-5
Dropout	0.3	0.3

최고 성능 (Dev 기준)

데이터셋	Dev Accuracy	Dev F1 Score
SST-5	51.8%	49.0%
CFIMDB	98.8%	98.8%

→ **SST-5**는 클래스 수가 많아 상대적으로 낮은 성능을 보임

→ **CFIMDB**는 단순한 이진 분류로 높은 정확도를 기록함.

→ **Full fine-tuning 방식**이 두 데이터셋에서 안정적으로 작동함.

GPT는 decoder-only 구조로,
입력 전체를 순차적으로 처리하며 문맥 정보를 누적하므로,
전체 파라미터를 조정하면 표현 흐름과 문맥 일관성이 강화됨.

ULMFiT 전략 사용

학습 방식 및 하이퍼파라미터

- GPT-2 모델에 ULMFiT 전략 적용
- Gradual Unfreezing + Discriminative LR + Slanted Triangular Scheduler
- Optimizer: AdamW
- 실험 목표: 학습 안정성과 파라미터 효율성 확보

항목	SST-5
Batch Size	64
Epochs	10
Learning Rate	1e-3
Dropout	0.3

최고 성능 (Dev 기준)

데이터셋	Dev Accuracy	Dev F1 Score
SST-5	50.9%	48.0%

Baseline vs. ULMFiT 성능 비교 (Dev 기준)

항목	Baseline (Full FT)	Task A (ULMFiT)
Accuracy	51.8%	50.9%
F1 Score	49.0%	48.0%

실험 결과 분석

1. GPT-2는 decoder-only 구조로, 표현 흐름의 연속성이 중요함
2. Gradual Unfreezing으로 레이어 간 흐름이 단절되어 성능 저하 발생함
3. 감정 분석은 정서적 미세 표현을 포착해야 하므로 전체 fine-tuning이 유리

결론 GPT-2 구조 특성상 ULMFiT보다 Full Fine-Tuning 방식이 더 안정적으로 작동

PART II – 패러프레이즈 탐지 실험

데이터셋 소개

Quora Question Pairs (QQP)

- 태스크: 두 문장이 의미적으로 유사한지 이진 분류 (0: 비유사, 1: 유사)
- 입력 포맷: "문장1 [SEP] 문장2" 형식으로 결합 후 GPT-2 tokenizer에 인코딩
- 출력 처리: [CLS] 없음 → 마지막 토큰의 hidden state를 추출 → Linear layer → Binary classification

BASELINE 학습

학습 방식 및 하이퍼파라미터

- 학습 방식: Full Fine-Tuning
- GPT-2 모든 파라미터 학습
- 마지막 토큰의 hidden state → Linear Layer → 2-class 분류
- Loss: CrossEntropyLoss
- Optimizer: AdamW

항목	값
Batch Size	8
Epochs	10
Learning Rate	1e-5
Dropout	미적용
환경	Colab L4 (A100), 총 약 471분 소요

최고 성능 (Dev 기준)

Accuracy	F1 Score	Precision	Recall
89.88%	89.19%	85.16%	87.83%

→ Baseline: GPT-2의 전체 파라미터를 학습하는 Full Fine-Tuning 방식으로 구성

→ Dropout 없이 학습되었음에도 불구하고 높은 Accuracy와 F1 Score를 달성.

→ 특히 Precision이 85.16%로 높아, false positive를 효과적으로 억제하는 데에 강점.

Task C 전략 사용

학습 방식 및 하이퍼파라미터

- GPT-2 모델에 복합 fine-tuning 전략 적용
- 0.7 × Mean + 0.3 × Last Hidden 혼합 Pooling
- Multi-Sample Dropout (3회 반복 평균)
- Label Smoothing 적용 CrossEntropyLoss (ε = 0.1)
- Optimizer: AdamW + LLRD
- (하위 레이어 0.5×, 상위 1×, head 2×)
- Scheduler: CosineAnnealingLR

항목	값
Batch Size	8
Epochs	10
Learning Rate	1e-5
Dropout	0.1~0.3

최고 성능 (Dev 기준, Epoch 6)

Accuracy	F1 Score	Precision	Recall
89.63%	88.99%	83.95%	88.83%

Baseline vs. Task C 성능 비교 (Dev 기준)

항목	Baseline (Full FT)	Task C
Accuracy	89.88%	89.63%
F1 Score	89.19%	88.99%
Precision	85.16%	83.95%
Recall	87.83%	88.83%

실험 결과 분석

항목	Baseline	Task C	비고
Accuracy	89.88%	89.63%	▼ 0.25%
F1 Score	89.19%	88.99%	▼ 0.20%
Precision	85.16%	83.95%	▲ 1.21%
Recall	87.83%	88.83%	▲ 1.00%

1. Accuracy/F1 점수는 비슷하나, Task C는 Recall에서 더 우수
2. Multi-Sample Dropout과 Label Smoothing → 과적합 방지 + 안정적 성능
3. Mean+Last Pooling → 문장 의미와 결론 정보 모두 활용해 표현력 향상
4. LLRD → 하위 표현 보존 + 상위 계층 태스크 적응
5. 학습 시간도 약 57분 단축 (471분 → 414분)

PART II – 시 생성 실험

데이터셋 소개

William Shakespeare's Sonnet

- 각 소네트는 14행으로 구성됨
- 전통적인 영국식 소네트 구조(ABAB CDCD EFEF GG)를 따름

BASELINE 학습

학습 방식 및 하이퍼파라미터

- 마지막 레이어 4개만 학습하는 Fine-Tuning 방식 사용
- CrossEntropyLoss 적용
- Optimizer: AdamW
- Batch Size: 8
- Epochs: 10
- Learning Rate: 1e-5

Generation 함수를 위한 전략

1.줄 수 제어

14줄 도달 시 자동 종료 → 소네트 형식 준수

2. 문장 길이 기반 구두점 유도

일정 길이 초과 시 마침표·쉼표 확률 증가

→ 문장 종결을 유도하여 자연스러운 흐름 유지

3.Top-p 샘플링

Top-p sampling: 누적 확률 기준으로 선택 후보 제한

→ 다양성, 품질의 균형 확보

성능 향상 기법

Unlikelihood Loss, Prefix-Tuning 기반 3가지 실험 계획

실험	기법	핵심 아이디어	기대 효과
Task A	Unlikelihood Loss	반복 토큰 확률 억제	반복 감소, 운율 개선
Task B	Prefix-Tuning	가상 prefix로 문체 조정	문체 제어, 창의성 향상
Task C	Unlikelihood Loss + Prefix-Tuning	반복 억제 + 문체 제어	자연스럽고 품질 높은 소네트

성능 향상 기법

실험 결과

모델	Perplexity	Distinct-1	Distinct-2	Rhyming Accuracy
Baseline	55.73	0.621	0.927	0.141
Task A	51.68	0.613	0.919	**0.215**
Task B	**50.11**	0.572	0.878	0.166
Task C	55.70	**0.641**	**0.944**	0.132

실험 결과 분석

- 제안한 방법들은 전반적으로 baseline보다 향상된 성능을 보였다.
- Perplexity는 Task A와 Task B에서 더 나은 언어 모델링 성능을 나타냈다.
- 어휘 다양성은 Task C가 가장 뛰어나, Distinct-1과 Distinct-2에서 baseline을 모두 초과했다.
- 운율 정확도는 Task A가 0.215로 가장 높았다
- Unlikelihood Loss를 적용한 Task A와 C는 학습 시간이 더 소요되었다.

- 요약하자면, Task A는 Perplexity와 운율에서 균형 잡힌 성능을 보였고, Task C는 어휘 다양성 면에서 가장 우수한 결과를 기록하였다.

결론 및 향후 실험

결론 및 향후 실험

프로젝트 결론

- GPT-2 기반 모델을 직접 구현하고,
  감정 분석, 패러프레이즈 탐지, 시 생성 세 가지 NLP 태스크에 적용
- 다양한 fine-tuning 전략을 비교하며 구조별 성능 차이와 적용 가능성 실증
- Full Fine-Tuning은 감정 분석에서 가장 안정적인 성능 달성
- ULMFiT, LoRA는 GPT-2 구조와 부적합하거나 성능 저하 유발
- 패러프레이즈 탐지에서는 Representation 조합, Label Smoothing, LLRD 기법 등으로 성능 향상
- 시 생성 태스크에서는 Prefix-Tuning + Unlikelihood Loss 조합이 창의성과 다양성 모두에 기여

향후 실험

결론 및 향후 실험

프로젝트 결론

향후 실험

- 감정 분석: encoder 구조 기반 모델(BERT 등)과의 비교 실험 확장 예정
- 패러프레이즈 탐지: 의미 불일치/반의문장 분류 등으로 확장
- 시 생성: 다양한 운율 구조 학습, 강화학습 기반 창의성 평가 도입 예정