자연어처리 2025 기말 프로젝트

컴퓨터공학과 2019112051

손민호

서론

본 프로젝트의 목적은 GPT-2 모델을 직접 구현하고, 이를 활용하여 감정 분석(Sentiment Analysis)과 문장 패러프레이즈 판별(Paraphrase Detection) 등 다양한 자연어처리(NLP) 태스크의 성능을 실험적으로 검증하는 것이다. 특히, 도메인과 클래스 수가 다른 여러 공개 데이터셋을 통해, 데이터 조합 및 도메인 차이가 모델 성능에 미치는 영향을 관찰하고자 하였다.

실험 내용

데이터셋 및 사전처리

SST(Stanford Sentiment Treebank):

영화 리뷰 데이터, 라벨이 0 ~ 4 (매우 부정 ~ 매우 긍정)까지 5 개 클래스

CFIMDB:

영화 리뷰 데이터, 0/1 의 2 개 클래스

Yelp Review Full:

음식점 리뷰 데이터, 별점 1 ~ 5(라벨 0 ~ 4)로 5 개 클래스.

전체 65 만여 개에서 1 만 개를 샘플링

모델 및 학습 전략

모델 구조:

직접 구현한 GPT-2 기반 분류기

- **실험 1:**

  SST, CFIMDB 각각으로 단독 학습 후, 두 데이터셋 상호 검증

- **실험 2:**

  SST 와 유사한 클래스 구조의 Yelp 데이터와 합쳐 학습 후,

  SST, CFIMDB 에 대해서 검증

- **실험 3:**

  Quora Paraphrase Detection 데이터셋을 사용해 Epoch 별

  패러프레이즈 판별 성능 실험

학습은 동일한 하이퍼파라미터로 진행하였고, 검증 정확도가 최고일 때의 모델 가중치를 저장하여 사용했다.

결과

Sentiment Analysis

학습|검증	sst	cfimdb
sst	0.478	0.588
cfimdb	0.550	0.878

SST 와 CFIMDB 는 라벨 체계가 달라 직접 비교가 어려워, SST 의 라벨을 0/1 로 이진화(0,1→0; 3,4→1)하여 범용성 비교 실험을 수행하였다. 단일 도메인 내에서는 CFIMDB 로 학습 시 더 높은 검증 정확도(0.878)가 나왔으나, 교차 검증에서는 SST 로 학습한 모델이 CFIMDB 에 대해 더 높은 성능을 보였다(0.588 vs. 0.550). 이는 SST 데이터가 중립 클래스(2) 등 더 다양한 문장 표현을 포함하여, 일반화 능력이 상대적으로 더 높았음을 보여준다.

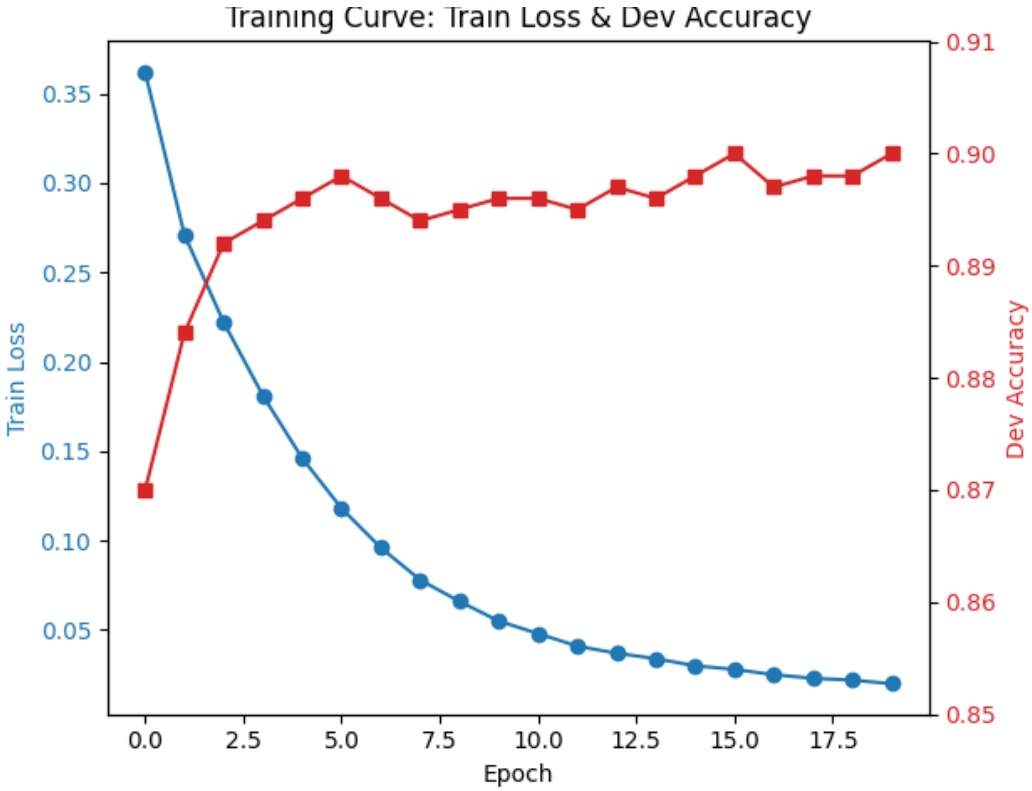
이를 반영하여 라벨 클래스의 종류가 5 개이고 SST 와 비슷한 데이터셋을 추가로 가져와 학습했을 때 범용성이 더 높아지는지 실험해 보았다.

Yelp 데이터셋이 65 만개 정도 되는데 이중에서 SST 와 비슷한 양인 1 만개를 추출한뒤 SST 데이터셋과 합쳐 약 2 만개의 데이터셋으로 구성한 뒤에 학습해 보았다. SST 에 대한 검증 정확도는 0.441, CFIMDB 에 대한 검증 정확도는 0.580 으로 SST 만 사용해서 학습했을 때보다 나쁜 결과를 얻었다. 이에 대한 이유로는 SST 는 영화 리뷰에 대한 데이터셋이고, Yelp 는 음식점 리뷰에 대한 별점 1~5 개 데이터셋이기 때문에 서로 분야가 달라서 학습에 안좋은 영향을 끼친것으로 분석된다.

Paraphrase Detection

Quora Paraphrase 데이터셋은 기본 설정인 10 Epoch 로는 모자랄 것으로 예상되어 20 Epoch 로 학습하였다.

SST 등 기존 데이터셋보다 약 28 만 개로 훨씬 커서, 학습에 소요되는 시간이 epoch 당 50 분 내외로 더 오래 걸렸다.



학습 결과를 그래프로 확인해보니 Epoch 16 과 20 에서 검증 정확도가 0.9 에 달하고, 로스가 충분히 작아진것으로 보아 학습은 epoch 20 정도로 충분하다고 볼 수 있다.

결론

본 프로젝트에서는 직접 구현한 GPT-2 기반 분류기를 활용하여 Sentiment Analysis, Paraphrase Detection 과제에 대해 실험을 수행하였다. 모델의 정확도 향상을 위해 SST-5, CFIMDB, Yelp 등 서로 도메인과 클래스 체계가 다른 다양한 데이터셋을 활용하여 학습을 진행하였으나, 실험 결과 데이터의 단순한 양적 증가가 반드시 모델 성능 향상으로 이어지지 않음을 확인할 수 있었다. 특히, 도메인 특성이나 감성 표현 방식(라벨 클래스)의 차이로 인해 서로 다른 데이터셋을 합칠 경우, 오히려 성능이 저하되는 현상이 관찰되었다. 이러한 결과는 자연어처리 모델의 성능을 높이기 위해서는 데이터의 양뿐만 아니라, 도메인에 대한 적절한 전처리 과정과 클래스 구조의 일관성을 함께 고려하는 것이 중요하다는 것을 알 수 있었다.