LoRA-PETM: Prediction Error Tracking을 위한 메모리 효율적 아키텍처

고욱[1], 김조아[2], 박수빈[3], 이연재[4], 최윤서[5], 안현주[6]
[1]서강대학교 기계공학과 학부생
[2]이화여자대학교 전자전기공학과 학부생
[3]숭실대학교 전자정보공학부 학부생
[4]숭실대학교 소프트웨어학부 학부생
[5]삼육대학교 컴퓨터공학부 학부생
[6]리안기술사사무소
nasakw@sogang.ac.kr, hellojoa1202@ewha.ac.kr, psb21@soongsil.ac.kr,
mycasseu@soongsil.ac.kr, younseo0611@syuniac.kr, suzic@nate.com

LoRA-PETM: A Memory-Efficient Architecture for Prediction Error Tracking

Wook Ko[1], Joa Kim[2], Subin Park[3], YeonJae Lee[4], YounSeo Choi[5], Hyun-Joo An[6]
[1]Dept. of Mechanical Engineering, Sogang University
[2]Dept. of Electronic and Electrical Engineering, Ewha Womans University
[3]Dept. of Electrical Engineering, Soogsil University
[4]Dept. of Software, Soongsili University
[5]Dept. of Computer Engineering, Samyook University
[6]LeeAhn Professional Engineer's Office

요 약

적대적 학습을 활용하는 예측 오류 추적 방법(PETM)의 메모리 비효율성 문제를 해결하기 위해, PCA 및 t-SNE 분석을 통해 두 모델 간 차이가 저차원 공간에 집중됨을 밝혔다. 이에 착안하여, 단일 모델에 저계수 근사(LoRA) 어댑터를 적용하고 지식 증류로 학습시켜 기존 PETM의 동작을 모방하는 LoRA-PETM 아키텍처를 제안한다. 실험 결과, 제안 모델은 절반의 파라미터로 기존 적대적 모델의 예측을 96% 재현하고 유사한 Recall(재현율) 성능을 보여, PETM을 효과적으로 경량화할 수 있음을 입증했다.

1. 서론

본 논문은 Prediction Error Tracking Method(PETM)[1]를 구성하는 두 모델의 가중치 공간을 분석하여, 적대적 학습에 기인하는 차이가 특정 저차원 부분 공간에 집중됨을 입증하였다. 이 발견에 착안하여, 저계수 근사(LoRA)[2]와 지식 증류를 결합함으로써 메모리 요구량을 크게 줄인 효율적인 아키텍처를 제안한다.

2. 관련 연구

2.1. 모델 아키텍처 및 학습

모델 A와 모델 A′는 모두 동일한 아키텍처로 구성한다. 모델 A′은 모델A에서 올바르게 예측하지 못한 이미지를 FGSM(Fast Gradient Sign Method)기법으로 생성한 적대적 예제로 교체한 데이터셋으로 학습한다[1].

3. 제안 이론

3.1 합성곱 계층의 주성분 분석

본 연구에서는 가장 큰 가중치 차이를 보인 Conv1 계층의 가중치 텐서 ($W \in \mathbb{R}^{32\times1\times3\times3}$)를 분석 대상으로 한다. 3×3 크기를 갖는 각 필터를 펼쳐 9차원의 특징 벡터(feature vector)로 만든다. 이에 따라 32개의 필터와 9개의 특징으로 구성된 행렬 $X \in \mathbb{R}^{32\times9}$ 를 구성한다. 이후 X의 공분산 행렬에 대해 주성분 분석(PCA)를 수행한다.

3.2 LoRA-PETM 아키텍처 설계

2.2의 분석에서 표준 모델과 A′ 모델의 가중치 차이가 저차원 부분 공간에 집중되어 있음을 확인했다. 이에 착안하여, LoRA 어답터를 큰 가중치 차이를 보이는 fc1, fc2, conv1 레이어에 적용하였다. 이를

통해 fc1(128×9216) 레이어의 경우 원본 행렬의 128 차원에 비해 절반(rank64)으로 근사하였다. LoRA-PETM 아키텍처에서 A, A′ 의 추론을 모두 구현하기 위해 파라미터 α를 도입하였다. α가 도입된 LoRA-PETM 아키텍처의 가중치는 아래 식과 같다. W_0 는 A 모델 각 레이어의 가중치, BA는 A′ 모델을 근사하는 저차원 행렬 LoRA 어답터 행렬이다[2]. α=0, α=1인 경우 각각 A, A′ 모델의 가중치에 해당한다.

$$W(\alpha) = W_0 + \alpha(BA) \text{(B, A는 Low rank 행렬)}$$

3.3 지식 증류를 활용한 LoRA 어댑터 학습

메모리 요구량을 줄여 효율적인 아키텍처를 구성하기 위하여, 지식 증류를 적용한다. LoRA 어답터가 (학생 모델)의 예측 로짓이 A′(교사모델)의 예측 로짓의 확률 분포를 모방하도록 Kullback-Leibler 발산을 손실함수로 하여 학습한다. 이를 통해 1.2M 파라미터 규모의 A′ 모델을 0.6M 파라미터 규모의 LoRA 어답터로 근사한다.

4. 실험 및 고찰 (Experiment & Discussion)

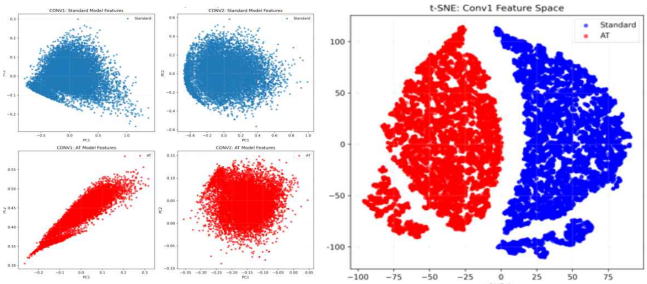
4.1 PETM의 매커니즘: 필터 공간의 구조적 분기

- A, A′ 모델의 Conv1, Conv2 PCA 분석

그림1 좌측의 PCA 분석은 두 모델의 Conv1/Conv2 필터 가중치의 PC1-PC2 평면 투영이다. 좌측 2개의 산포도는 Conv1 계층, 우측은 Conv2 계층을 나타낸다. A의 가중치(푸른색)가 분산된 것과 반대로, A′의 가중치(붉은색)는 Conv1 계층에서 가중치가 소수의 주성분 축으로 집중되어 단순화된 저차원 구조를 형성한다.

- 특징 공간의 위상학적 이분화

그림1 우측의 Conv1 계층의 특징 공간 t-SNE 에서 A와 A′의 특징 벡터가 두 개의 뚜렷하게 분리된 군집을 형성함을 확인했다. 이는 두 모델이 근본적으로 다른 내부 표현을 학습함을 의미한다. 이는 A′의 예측이 A와 서로 다른 방식을 통한 검증으로 작용함을 의미한다.



(그림1) A 모델과 A‘ 모델 특성 공간의 시각화

4.2 LoRA-PETM 아키텍처 구현

아래 표1은 LoRA-PETM 아키텍처와 기존 A′ 모델의 예측 결과 일치도를 나타낸다. 200 에포크의 학습을 통해 A’ 모델과 0.96의 일치도를 나타낸다.

(표1) LoRA-PETM 과 기존 모델의 예측 결과 비교

구분	LoRA-PETM & A	LoRA-PETM & A′
일치도	1.0	0.96

표2의 결과는 A 모델의 예측 결과에 낮은 신뢰도를 부여하는 경우가 증가하여 Recall이 소폭 증가하고, 정밀도가 소폭 하락하였다. 메모리 요구량을 개선하기 위해 A’ 모델의 파라미터 규모를 50%로 줄였음에도, LoRA-PETM 아키텍처는 기존 모델과 동등한 수준의 성능을 유지하였다.

(표2) LoRA-PETM 과 기존 모델의 성능 비교

구분	PETM	LoRA-PETM
재현율(Recall)	87.50%	88.42%
정밀도(Precision)	59.04%	53.48%

5. 결론 (Conclusion)

예측 오류 탐지 과업의 어려움[1]을 고려할 때, 기존에 비해 75%의 메모리만을 사용하여 실용적인 수준의 신뢰도 평가 성능을 LoRA-PETM이 메모리 효율적으로 구현함을 시사한다. 추후 해당 아키텍처의 스케일링하여 대규모 모델에 적용하는 연구가 요구된다.

ACKNOWLEDGEMENT

※ 본 논문은 과학기술정보통신부 대학디지털교육역량강화 사업의 지원을 통해 수행한 ICT 멘토링 프로젝트 결과물입니다.

참고문헌

[1]Lee, G., Yun, C., Kim, S., & Kang, B. B., Discriminator to grade classification results of neural networks via Prediction Error Tracking Method, Knowledge-Based Systems, 310, 112954, 2025.

[2] Hu, E. J., et al., LoRA: Low-Rank Adaptation of Large Language Models, International Conference on Learning Representations (ICLR), 2022.