LoRA 기반 파인튜닝 모델의 오염 학습 샘플 사후 순위화 기법

강병수*, 김민석**, 구형준**

*명지대학교 (학부생), **성균관대학교 (대학원생, 교수)

Post-hoc Ranking of Poisoned Training Samples in LoRA-Fine-Tuned Models

Byungsoo Kang*, Minseok Kim**, Hyungjoon Koo**,

*Myongji University(Undergraduate student)
**SungKyunkwan University(Graduate student, Professor)

요 약

데이터 오염이 의심되는 모델을 조사하려면 학습에 사용된 표본 중 우선 검토할 대상을 좁혀야 한다. 본 연구는 LoRA 어댑터의 가중치 변화와 학습 표본의 은닉표현을 이용해 오염 의심 순위를 정하는 사후 분석 방법을 제안한다. 가중치 변화에 특이값 분해를 적용해 주요 방향을 구하고, 응답 생성 이전의 명령문 은닉표현을 이 방향에 투영하여 각 표본의 점수를 계산한다. 추가 학습이나 역전파 없이 표본당 한 번의 순전파로 채점할 수 있다. Qwen3-1.7B와 BackdoorLLM 자료로 구성한 18개 설정에서 SS, SPECTRE, LoRAScan 재구현과 비교하였다. 제안 방법은 평균 AUROC 0.857과 평균정밀도(AP) 0.498을 기록하여 비교 방법 중 가장 높은 평균 성능을 보였다. 이러한 결과는 제안 방법이 오염 의심 표본의 검토 순서를 정하는 데 활용될 수 있음을 보여준다.

I. 서론

거대 언어 모델 (Large Language Model, LLM)을 특정 도메인과 태스크에 효율적으로 적응 시키는 방법으로 저순위 적응 (Low-Rank Adaptation, LoRA) 기법을 [1] 널리 활용한다. LoRA는 원본 모델의 파라미터를 고정한 채 일부 저차원 파라미터만을 학습하므로 비교적 적은 비용으로 파인튜닝을 할 수 있다는 장점이 있다. 그러나 실제 파인튜닝 과정에서는 데이터의 출처와 신뢰성을 완전히 보장하기 어려운 경우가 많으며, 소수의 악성 샘플이 삽입되는 데이터 오염 공격에 노출될 수 있다.

이러한 위협에 대응하기 위해 기존 연구는 주로 배포 전 어댑터의 오염 여부를 판정하거나 [2, 3] 추론 시점에서 트리거 입력을 탐지 [4]하는 방법에 초점을 맞췄다. 그러나 모델의 이상 동작이 실제로 확인된 이후에는 모델 자체의 오염 여부를 판단하는 것 뿐만이 아니라, 어떤 학습 데이터가 해당 행위를 유발했는지 추적하는 과정이 필요하다. 사고 발생 후 모델과 학습 데이터가 확보된 상황에서도, 모델의 악성 동작을 유발한 학습 샘플을 역추적하는 문제는 충분히 다뤄지지 않았다.

본 연구는 이러한 문제를 해결하기 위해 LoRA 가중치 변화와 학습 샘플의 은닉 표현을 결합한 사후 오염 샘플 순위화 방법을 제안한다. 핵심 아이디어는 파인튜닝 과정에서 형성된 LoRA 가중치 변화의 주요 방향을 특이값 분해 (Singular Value Decomposition, SVD)를 통해 추출하고, 각 학습 샘플의 은닉 상태가 해당 방향과 얼마나 정렬되는지를 이용하여 오염 의심도를 정하는 것이다. 이를 통해 각 학습 샘플에 점수를 부여하고, 분석가가 오염 가능성이 높은 샘플을 우선적으로 조사할 수 있도록 순위를 제공한다.

본 논문의 주요 기여점은 다음과 같다.

첫째, LoRA 어댑터의 가중치 변화에서 추출한 특이벡터와 학습 샘플의 은닉 표현을 이용하여 학습 샘

* 이 논문은 2026년도 정부(경찰청)의 재원으로 과학치안진흥센터 사이버수사지원기술개발 사업의 지원을 받아 수행된 연구임 (RS-2026-25541010).

플의 오염 의심도를 순위화하는 분석 방법을 제안한다.

둘째, 오염 샘플에 대한 사전 라벨이나 트리거 정보 없이도 분석 방향의 부호와 대상 층을 결정하는 규칙을 설계하고, 다양한 오염 환경에서 제안 방법의 순위화 성능을 평가한다.

II. 배경 및 관련 연구

저순위 적응. LoRA는 기존 모델의 가중치를 고정한 상태에서 대상 선형 모듈에 추가되는 저순위 행렬을 추가하고, 해당 파라미터만을 학습하는 파라미터 효율적 파인튜닝 기법이다 [1]. 학습된 저순위 가중치는 원본 모델과 분리된 어댑터로 저장될 수 있으므로, 파인튜닝 과정에서 형성된 가중치 변화를 독립적인 분석 대상으로 확보할 수 있다.

데이터 오염 공격. 데이터 오염 공격은 학습 또는 파인튜닝 데이터셋에 소수의 악성 샘플을 삽입하여, 모델이 공격자가 의도한 동작을 하도록 유도하는 공격이다. 오염된 모델은 일반적인 입력에서는 정상적으로 작동하는 것처럼 보이지만, 특정 트리거나 입력 조건이 만족되면 공격자가 지정한 비정상적인 동작을 수행할 수 있다. 따라서, 모델의 출력만으로는 학습 과정에서의 오염 여부를 식별하기 어렵다.

기존 백도어 탐지 연구. 기존 백도어 탐지 연구는 주로 학습된 어댑터 자체의 악성 여부를 판정하거나 [2, 3], 추론 시점의 입력에서 트리거가 포함된 입력을 탐지하는 데 초점을 둔다 [4]. 이러한 접근은 어댑터 또는 입력을 분석 대상으로 하며, 개별 학습 샘플에서 원인이 된 샘플의 오염 가능성을 평가하지 않는다. 학습 표본 단위의 탐지 방법으로는 전체 학습데이터의 은닉표현으로부터 주요 이상 방향을 이용하는 Spectral Signatures [5]와 오염 표본의 영향을 줄여 데이터 분포를 추정하고, 이를 바탕으로 의심 표본을 찾는 SPECTRE [6]가 있다. 반면 본 연구는 학습된 LoRA 어댑터의 가중치 변화에서 분석 방향을 먼저 추출하고 각 학습 샘플을 독립적으로 채점하므로, 데이터셋 전체의 활성화 행렬을 구성하거나 샘플별 역전파를 수행하지 않는다.

III. 제안 방법

그림 1은 제안 방법의 전체 과정을 나타낸다. 본 방법은 먼저 학습된 LoRA 어댑터의 가중치 변화로부터 분석 방향과 대상 층을 결정한 뒤, 이후 각 샘플을 모델에 입력하여 선택된 층의 은닉 상태를 추출하고, 해당 은닉 상태와 분석 방향의 내적을 이용해 오염 의심 점수를 계산한다. 최종적으로 각 샘플을 점수에 따라 정렬하여 오염 가능성이 높은 샘플을 우선적으로 제시한다.

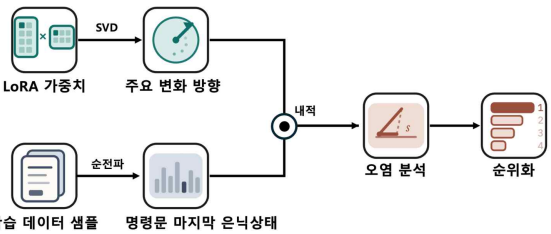


그림 1 제안 방법의 전체 흐름. LoRA 어댑터에서 분석 방향과 대상 층을 결정하고, 각 학습 샘플의 은닉 상태와 분석 방향의 내적을 이용하여 오염 의심 점수를 산출한다.

3.1 위협 모델 및 분석 전제

본 연구는 데이터 오염 가능성이 있는 학습 데이터로 LoRA 파인튜닝 이후, 학습된 모델과 해당 학습 데이터셋을 확보하여 분석하는 상황을 가정한다. 분석가는 개별 샘플의 오염 여부와 전체 오염률, 그리고 트리거의 구체적인 형태를 사전에 알지 못한다. 다만 오염 샘플은 전체 학습 데이터 절반 이하라고 가정한다. 분석의 목적은 별도의 오염 라벨이나 트리거 정보 없이 각 샘플의 오염 의심 점수를 산출하고, 이를 바탕으로 우선적으로 검토할 샘플을 순위화하는 것이다.

3.2 방향 벡터 추출 및 샘플 점수 산출

본 연구는 LoRA 어댑터 내 다층 퍼셉트론의 W_{down} 모듈에 적용된 LoRA 가중치를 분석한다. 구체적으로, 각 층의 LoRA 가중치 변화 행렬에 SVD를 적용하고, 가장 큰 특이값에 대응하는 제1특이벡터 $u_1^{(l)}$을 구한다. 여기서 $u_1^{(l)}$은 해당 층에서 LoRA에 의해 형성된 가중치 변화가 가장 크게 나타난 축을 나타내며, 이후 학습 샘플의 은닉 표현을 평가하기 위한 분석 방향으로 사용한다. LoRA가 적용된 여러 층 가운데 실제 분석에 사용될 층은 가장 큰 제1특이값을 갖는 층으로 결정한다. 즉, 분석 층 l^*은 다음과 같이 선택한다.

$$l^* = \arg\max_l \sigma_1^{(l)} \quad (1)$$

방법	AUROC	AP	P@25	R@25
LoRAScan [4]	0.531	0.093	0.087	0.057
Spectral Signatures [5]	0.724	0.411	0.418	0.319
SPECTRE [6]	0.764	0.434	0.398	0.301
제안 방법	**0.857**	**0.498**	**0.467**	**0.444**

표 1 오염 학습 표본의 순위화 성능 비교. 트리거 6종과 오염률 3종 (2%, 5%, 10%)을 조합한 18개 설정에서 각각 학습 표본 500개를 순위화하고, 설정별 성능을 동일 가중치로 평균하였다.

수식 1에 따라 선택된 층의 $u_1^{(l^*)}$을 최종 분석 방향으로 사용한다. σ_1은 각 층의 LoRA 가중치 변화에서 가장 큰 성분의 크기를 나타내며, 이 과정에서는 오염 라벨이나 은닉 상태를 사용하지 않고 어댑터 가중치만으로 분석 층과 방향을 결정할 수 있다.

3.3 부호 결정과 층 선정

분석 층과 방향을 결정한 뒤, 각 학습 샘플을 모델에 입력하여 선택된 층에서 명령문 마지막 토큰의 은닉 상태를 추출한다. 이후 비정상적으로 큰 활성값을 갖는 일부 차원을 마스킹하고, 전처리된 은닉 상태와 $u_1^{(l^*)}$의 내적값을 해당 샘플의 오염 의심 점수로 사용한다. 한편 SVD로 얻은 $u_1^{(l^*)}$의 부호는 임의적으로 결정되므로, 동일한 분석 축에서도 어느 방향이 높은 오염 의심 점수에 대응하는지는 결정되어 있지 않다. 본 연구에서는 오염 샘플이 전체 데이터의 절반 이하라는 가정에 따라 점수 분포의 다수를 정상 샘플로 간주한다. 점수 하위 80%를 기준 분포로 결정하고, 해당 분포의 표준편차를 기준으로 다수 방향에 위치하도록 $u_1^{(l^*)}$의 부호를 결정한다. 최종적으로 부호가 보정된 점수를 기준으로 학습 샘플을 내림차순 정렬하여 오염 의심 순위를 산출한다.

IV.실험 및 결과 분석

4.1 실험 설정

본 연구는 Qwen3-1.7B 모델에 LoRA (r=16, α=32)를 적용하고, 정상 및 오염 표본을 섞은 500개의 데이터로 각 모델을 4 epoch 학습하였다. BackdoorLLM [7]의 negsentiment 데이터를

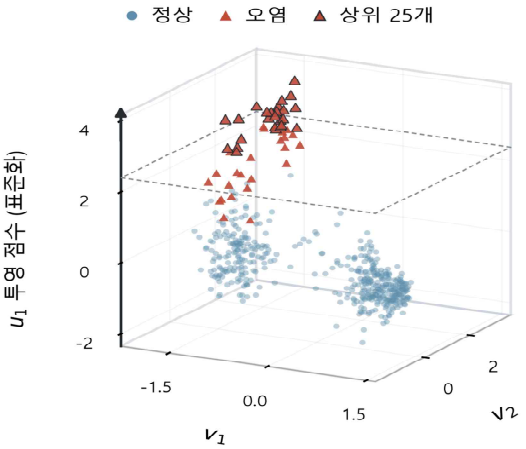


그림 2 $u_1^{(l^*)}$투영 점수에 따른 정상·오염 학습 표본의 분포 (badnet, 오염률 10%). 파란 원 450개와 붉은 삼각형 50개는 각각 정상·오염 표본이다. 세로축의 투영 점수가 높을수록 먼저 검토 우선순위가 높다.

바탕으로 트리거 6종 (badnet, syntax, mtba, word, sleeper, phrase)과 오염률 3종 (2%, 5%, 10%)을 조합한 18개 설정을 구성하였다. 각 설정의 오염 표본 수는 각각 10개, 25개, 50개이다.

각 설정에서는 학습에서 사용된 500개 표본 전체를 대상으로 오염 의심 순위를 산출하였다. 순위화에는 응답을 제외한 명령문 입력을 사용하였으며, 실제 오염 여부는 평가에만 활용하였다. 비교 방법은 Spectral Signatures (SS) [5], SPECTRE [6], LoRAScan [4]을 사용하였다. SS와 SPECTRE에는 제안된 방법론과 동일한 층 및 전처리를 적용한 마지막 입력 토큰 은닉표현을 제공하였으며, 별도의 클래스 구분 없이 전체 학습 표본을 분석하였다. SPECTRE의 예상 오염률은 실제 비율과 무관하게 10%로 고정하였다. LoRAScan은 명령문 내 토큰에 따른 LoRA 내부 활성값의 급증을 이용하며, 외부 정상 명령문 40개로 기준을 보정하여 재구현하였다.

순위화 성능은 총 네가지 지표로 평가하였다. AUROC는 오염 표본이 정상 표본보다 높은 순위에 위치하는 정도를, 평균 정밀도 (AP)는 순위 상단에 오염 표본이 집중되는지를 평가한다. 또한 분석가가 500개 중 25개 (5%)만 검토하는 상황을 가정하여, 검토한 25개 중 오염 표본의 비율을 Precision@25 (P@25), 전체 오염 표본 중 상위 25개에서 발견한 비율을 Recall@25 (R@25)로 측정하였다.

4.2 순위화 성능 비교

표 1은 트리거 6종과 오염률 3종을 조합한 18개 설정에서의 순위화 성능을 동일 가중치로 평균한 결과이다. 제안 방법은 AUROC 0.857, AP 0.498, P@25 0.467, R@25 0.444를 기록하여 모든 평가 지표에서 비교 방법보다 높은 성능을 보였다. AUROC에서는 가장 높은 비교 방법인 SPECTRE의 0.764보다 0.093 높은 성능을 보였으며, AP 역시 0.064 더 높았다. 또한, P@25와 R@25 또한 0.467과 0.444로 가장 높게 나타났다. 따라서, 제한된 수의 표본만을 검토하는 상황에서도 오염 표본을 우선순위화 하는데 효과적임을 확인하였다.

4.3 분석 방향과 은닉 표현의 관계

제안 방법이 은닉 표현의 어떤 차이를 순위화에 반영하는지 확인하기 위해 분석 방향과 학습 표본 분포의 관계를 분석하였다. SS가 은닉표현의 최대 분산 방향을 사용하는 것과 달리, 제안 방법은 LoRA 가중치 변화의 주요 방향인 $u_1^{(l^*)}$을 사용한다. 두 방향의 절대 코사인 유사도는 18개 설정에서 평균 0.061로 낮게 나타났다. 이는 제안 방법의 분석 방향이 은닉 표현의 단순한 최대 분산 방향과는 다름을 보여준다.

그림 2는 badnet 10% 설정에서 $u_1^{(l^*)}$ 투영 점수와 정상 및 오염 표본의 분포를 시각화한 결과이다. 세로축은 $u_1^{(l^*)}$ 투영 점수이고, 나머지 두 축은 $u_1^{(l^*)}$에 직교하는 부분공간의 분포를 나타낸다. 정상 표본이 직교 부분공간에서 여러 영역에 분포하는 가운데, 오염 표본은 $u_1^{(l^*)}$ 점수가 높은 영역에 집중되는 양상을 보였다. 따라서, 제안 방법은 직교 부분공간에서의 표본 위치와 관계없이 $u_1^{(l^*)}$ 점수가 높은 표본을 우선적으로 순위화한다. 해당 설정에서 상위 25개는 모두 오염 표본으로 나타나였으며, 전체 오염 표본의 50%를 발견하였다 (P@25=1.0, R@25=0.5). 또한, 전체 순위의 AP는 0.961이었다. 이러한 결과는 가중치에서 추출한 분석 방향과 오염 표본의 은닉 표현 차이가 잘 정렬되어 있으며, 이에 따라 높은 순위화 성능으로 이어짐을 보여준다.

V. 한계점

본 연구는 파인튜닝에 사용된 학습 데이터셋를 대상으로 오염 의심 순위를 산출하는데 초점을 두므로, 학습에 포함되지 않은 새로운 입력의 오염 여부를 판별하는 용도로는 적합하지 않다. 실제로, 미학습 데이터 200개 대상으로 평가한 결과 평균 AUROC는 0.626으로, 학습 표본을 대상으로 한 순위화 성능보다 낮게 나타났다. 또한 본 방법은 학습 표본의 상대적인 오염 의심 순위를 제공하지만, 실제 오염 샘플 수나 오염 여부를 구분하는 임계값을 자동으로는 결정하지는 않는다. 따라서 실제 활용 시에는 분석가가 상위 N개 의심 표본을 우선 검토하는 방식이 적합하며, 향후에는 미관측 입력으로의 일반화와 비지도 임계값 결정 방법에 대한 추가 연구가 필요하다.

VI. 결론

본 논문은 LoRA 어댑터의 가중치 변화와 은닉 표현을 이용하여 파인튜닝 데이터셋 내 오염 의심 표본을 순위화하는 분석 방법을 제안한다. 제안 방법은 BackdoorLLM 기반 18개 설정에서 기존 연구들에 비해 가장 높은 성능을 보였다. 이를 통해 LoRA 가중치와 학습 표본의 은닉 표현을 함께 활용하는 것이 오염 가능성이 높은 학습 표본을 우선적으로 식별하는데 효과적임을 확인하였다.

[참고문헌]

[1] Hu, E. J., et al. (2021). LoRA: Low-Rank Adaptation of Large Language Models. arXiv preprint arXiv:2106.09685.

[2] Merenciano, D. P., et al. (2026). Weight Space Detection of Backdoors in LoRA Adapters. arXiv preprint arXiv:2602.15195.

[3] Lelle, T. (2026). Token-Level Generalization in LoRA Adapter Backdoors: Attack Characterization and Behavioral Detection. arXiv preprint arXiv:2605.30189.

[4] Obidov, D., et al. (2026). LoRAScan: Detecting Backdoor Prompts in Low-Rank Adapters for Large Language Models via Down-Projection Activation Spikes. arXiv preprint arXiv:2608.06795.

[5] Tran, B., et al. (2018). Spectral Signatures in Backdoor Attacks. Advances in Neural Information Processing Systems, 31.

[6] Hayase, J., et al. (2021). SPECTRE: Defending against Backdoor Attacks Using Robust Statistics. In International Conference on Machine Learning (pp. 4129 - 4139). PMLR.

[7] Li, Y., et al. (2026). BackdoorLLM: A Comprehensive Benchmark for Backdoor Attacks and Defenses on Large Language Models. Advances in Neural Information Processing Systems, 38.