LLM을 활용한 텍스트 모델 인버전 결과의 유창성 복원 및 클래스 정보 보존 분석

김진성*, 최석환†

*연세대학교 (대학원생), †연세대학교 (지도교수)

LLM-based Fluency Restoration and Class Preservation Analysis for Text Model Inversion

Jin-Seong Kim* and Seok-Hwan Choi†

*Yonsei University(Graduate student), †Yonsei University(Professor)

요 약

텍스트 모델에 대한 인버전 공격은 이산적인 토큰 공간의 특성으로 인해 문법적으로 붕괴된 비유창한 토큰 나열을 생성하는 한계를 가진다. 본 논문에서는 이러한 인버전 결과에 잔여된 클래스 정보를 분석하고, 대규모 언어모델(LLM)을 활용한 사후 처리를 통해 유창한 문장으로 재구성할 수 있는지를 검토한다.실험 결과, 비유창한 토큰 시퀀스에도 상당한 클래스 판별 정보가 포함되어 있으며, LLM 기반 재구성을 통해 유창성을 크게 향상시키면서도 클래스 정보를 효과적으로 보존할 수 있음을 확인하였다.

I. 서론

딥러닝 기술의 발전으로 다양한 분야에서 성능 향상이 이루어졌으나, 모델이 학습 데이터를 암기하는 특성으로 인해 개인정보 유출 위험이 강력한 보안 문제로 대두되고 있다 [1].

특히 학습된 모델로부터 학습 데이터를 복원하는 모델 인버전(Model Inversion) 공격은 민감 정보 탈취 가능성을 보여주는 대표적인 프라이버시 위협이다 [2]. 기존 모델 인버전 공격은 주로 이미지 데이터를 대상으로 발전해 왔다. 텍스트 모델에 동일한 방식을 적용할 경우, 이미지는 연속적인 픽셀 공간에서 최적화가 가능한 반면 텍스트는 이산적인 토큰 공간을 다뤄야 한다는 본질적인 차이가 존재한다 [3].

이로 인해 기존의 텍스트 대상 공격은 이산 공간의 제약을 우회하기 위해 연속적인 임베딩 공간이나 로짓 공간에서 최적화를 수행한 뒤 가장 가까운 토큰으로 매핑하는 방식을 주로 사용한다. 그러나 이러한 접근법들은 개별 토큰의 복원 성능에만 집중할 뿐, 매핑 과정에서 발생하는 정보 손실로 인해 원본 문맥이 파괴된 비유창한 토큰 나열이 생성되는 한계를 지닌다. 즉, 복원된 결과가 문법적으로 자연스러운 문장을 형성하는지에 대한 분석은 상대적으로 부족한 실정이다.

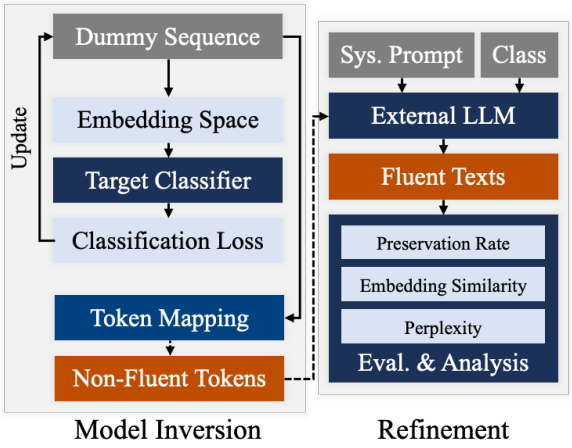

그림 1 제안하는 방법론의 개요도

표 1 비유창 토큰 시퀀스와 LLM을 통해 재구성된 문장의 유창성 및 임베딩 유사도 비교

Model	Dataset	Type	PPL	Sim.	Representative Example
BERT	AG News	Raw	11508.5	0.862	democrat eyeing troop bethlehem mur...
		Refined	581.2	0.848	Nationalist lawyer introduces immigrat...
	Emotion	Raw	11598.8	0.915	resignation regret hospital abortion sib...
		Refined	543.1	0.867	Sibling resignation brought residual re...
RoBERTa	AG News	Raw	344466.1	0.996	Hamas Assad Beirut Mosque Greece ...
		Refined	1171.3	0.997	Hamas and Assad condemned attack ...
	Emotion	Raw	370763.7	0.995	sorry dreadful poor numb witness ...
		Refined	1464.1	0.994	Poor numb witness endured dreadful ...

본 논문에서는 이를 완화하기 위해, 모델 인버전으로 얻은 비유창한 토큰 시퀀스를 대규모 언어모델(LLM)을 활용하여 자연스러운 문장으로 재구성하는 사후 처리 기반의 정보 유출 방법론을 제안한다.

II. 제안 방법

본 논문에서는 모델 인버전 공격으로 생성된 비유창한 토큰 시퀀스를 대규모 언어모델을 활용하여 자연스러운 문장으로 재구성하는 사후 처리 기반 방법을 제안한다.

2.1 비유창 토큰 시퀀스 획득

텍스트 분류 모델에 대한 모델 인버전 공격을 수행하여 각 클래스에 해당하는 클래스의 대표 정보를 갖고 있는 토큰 시퀀스를 획득한다. 본 연구에서는 연속적인 로짓 공간에서 최적화를 수행한 뒤 가장 가까운 토큰으로 매핑하는 방식을 사용한다. 이 과정에서 문법적 유창성 제약은 고려하지 않으며, 그 결과 문맥적 자연스러움이 파괴된 토큰 나열이 생성된다.

2.2 LLM 기반 문장 재구성

획득한 비유창 토큰 시퀀스를 대규모 언어모델에 입력하여 자연스러운 문장으로 재구성한다. 이때 두 가지 조건을 비교한다.

- 토큰 시퀀스 없이 목표 클래스 정보만을 제공. 이는 토큰에 잔여된 정보의 실제 기여도를 확인하기 위한 대조군으로 사용.
- 복원된 토큰 시퀀스와 클래스 정보를 함께 제공. 주어진 토큰을 활용하고, 해당 클래스에 적합한 유창한 문장을 생성하도록 유도.

이 두 조건을 통해, 단순히 클래스 정보만으로 생성한 문장과 인버전 토큰을 활용한 문장 사이에 클래스 정보 보존 및 의미적 유사도 측면의 차이를 면밀하게 분석한다.

2.3 클래스 정보 잔여도 평가 및 분석

재구성된 문장이 원래 목표 클래스의 정보를 얼마나 유지하고 있는지를 정량적으로 평가한다. 주요 평가 지표는 재구성된 문장이 원래 목표 클래스로 분류되는 비율인 Class Preservation Rate, GPT-2 모델을 통해 재구성된 문장의 유창성을 평가하는 Perplexity, 재구성된 문장과 원본 학습 데이터 간의 의미적 유사도인 Embedding Similarity로 구성된다.

결론적으로, 본 연구에서는 이러한 지표들을 통해 비유창한 토큰 시퀀스에 잔여된 클래스 정보를 유지함과 동시에, LLM 기반 사후 처리가 이를 얼마나 효과적으로 유창한 문장으로의 복원 가능성을 종합적으로 분석한다.

III. 실험 결과

본 실험에서는 범용적인 BERT와 RoBERTa를 타겟 모델로 사용하고, AG News와 Emotion 데이터셋에 대한 결과를 평가하였다.

3.1 유창성 의미 유사도 분석

표 1은 인버전으로 얻은 Raw 토큰 시퀀스와 LLM으로 재구성한 Refined 문장의 Perplexity와 임베딩 유사도를 비교한 결과이다. Raw 상태의 Perplexity는 매우 높게 나타나 문법적으로 붕괴된 상태임을 보여주었으나, Refined 이

표 2 LLM에 의해 생성된 문장의 클래스 정보 보존 성능 비교 결과

Data	Model	Accuracy (%)		Precision		Recall		F1-Score	
		w/o	w	w/o	w	w/o	w	w/o	w
AG News	BERT	77.00	97.50	0.865	0.975	0.770	0.975	0.770	0.975
	RoBERTa	79.25	98.75	0.827	0.988	0.793	0.988	0.792	0.988
Emotion	BERT	68.83	81.00	0.735	0.833	0.688	0.810	0.665	0.812
	RoBERTa	87.00	97.83	0.884	0.978	0.870	0.978	0.872	0.978

후 수치가 크게 감소하여 유창성이 개선되었음을 확인하였다. 한편 의미 유사도는 Raw와 Refined 모두 비교적 높은 수준을 유지하였으며, BERT에서는 Raw가 근소하게 높은 경향을 보였다. 이는 인버전 토큰이 클래스와 관련된 핵심 신호를 이미 포함하고 있음을 시사한다.

3.2 클래스 정보 보존 성능

표 2는 토큰 없이 클래스 정보만으로 생성한 문장(w/o)과 인버전 토큰을 함께 제공하여 재구성한 문장(w)의 클래스 보존 성능을 비교한 결과이다. 두 데이터셋 모두에서 토큰을 활용한 경우 정확도, Precision, Recall, F1-Score가 크게 향상되었다. 특히 AG News에서는 BERT와 RoBERTa 모두 97% 이상의 높은 보존률을 보였으며, Emotion에서도 뚜렷한 성능 향상이 관찰되었다. 이는 앞선 유사도 분석 결과와 일관되게, 비유창한 토큰 시퀀스 안에 유의미한 클래스 판별 정보가 잔여하고 있음을 보여준다.

3.3 종합 분석

실험 결과를 종합하면, 텍스트 모델에 대한 인버전 공격은 단순히 무의미한 토큰 나열을 생성하는 것이 아니라, 클래스 판별에 활용 가능한 실질적인 정보를 상당 부분 복원하고 있음을 확인할 수 있다. Raw 토큰은 문법적으로 붕괴되어 있음에도 불구하고 높은 의미 유사도를 보였으며, 이를 조건으로 제공했을 때 클래스 보존 성능이 대조군 대비 크게 향상되었다.

IV. 결론

본 논문에서는 텍스트 모델에 대한 인버전 공격으로 생성된 비유창한 토큰 시퀀스에 잔여된 클래스 정보를 분석하고, 대규모 언어모델을 활용한 사후 처리를 통해 이를 유창한 문장으로의 재구성 가능성을 실험적으로 검토하였다. 이러한 결과는 텍스트 모델 인버전 공격의 산출물이 실질적인 클래스 정보를 포함하고 있으며, 언어모델과의 결합을 통해 그 위협이 더 가시적인 형태로 발현될 수 있음을 시사한다.

[Acknowledgement]

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원(RS-2026-25472714) 및 과학기술정보통신부 및 정보통신기획평가원의 2026년도 의료AI반도체 전·후방 엔지니어 육성사업의 지원(2024-0-0096)을 받아 수행된 연구임.

[참고문헌]

[1] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos and D. Song, The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks, in Proc. of the 28th USENIX Security Symposium, 2019.

[2] M. Fredrikson, S. Jha and T. Ristenpart, Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures, in Proc. of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 1322-1333, 2015.

[3] H. Fang, Y. Qiu, H. Yu, W. Yu, J. Kong, B. Chong, B. Chen, X. Wang, S.-T. Xia and K. Xu, Privacy Leakage on DNNs: A Survey of Model Inversion Attacks and Defenses, arXiv preprint 2402.04013, 2024.