NEAT: 대형 오디오 언어 모델에서의 오디오-텍스트 정렬을 위한 Norm Equalization

김선욱, 박재홍, 강주연, 전용현, 홍지민, 김남수
서울대학교 전기정보공학부 뉴미디어통신공동연구소
{sukim, jhpark, jykang, yhjeon, jmhong}@hi.snu.ac.kr, nkim@snu.ac.kr

NEAT: Norm Equalization for Audio-Text Alignment in Large Audio Language Models

Seonuk Kim, Jaehong Park, Ju Yeon Kang, Yonghyeon Jun, Ji Min Hong, Nam Soo Kim
Department of Electrical and Computer Engineering and INMC, Seoul National University

Abstract

Large Audio Language Models (LALMs) extend LLMs to audio understanding by projecting audio embeddings into the LLM input space. We find that audio and text embeddings exhibit a severe L2 norm imbalance, and show that this mismatch leads to suboptimal audio-text alignment. To address this, we introduce NEAT (Norm Equalization for Audio-Text Alignment), which applies per-token L2 norm clipping to audio embeddings at inference time. On the MMAU benchmark using Audio Flamingo 3, NEAT improves overall accuracy from 74.4% to 76.1% without any additional training.

I. 서론

트랜스포머(Transformer)[1] 아키텍처가 등장하며, 대형 언어 모델(Large Language Model, LLM)은 언어 이해 및 생성 능력에서 비약적 발전을 이루었다. LLM의 정보 처리 능력을 텍스트 이외의 모달리티로 확장하려는 연구가 진행되고 있고, 특히 대형 오디오 언어 모델(Large Audio Language Model, LALM)은 오디오 인코더를 통해 음성, 환경음, 음악 등의 신호를 임베딩 벡터로 변환한 뒤, 모달리티 프로젝터를 통해 LLM의 임베딩 공간에 매핑함으로써 LLM에 오디오 이해 및 추론 능력을 부여한다.

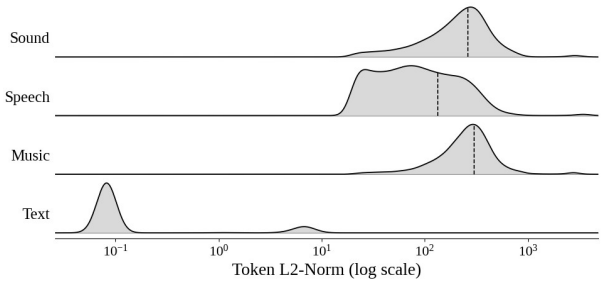

그림 1. Embedding Norm Distribution by Modality

이러한 구조는 오디오 신호로부터 의미론적 정보를 추출하고 이를 언어적 맥락과 함께 다룰 수 있다는 점에서 잠재력을 지니지만, 동시에 이질적인 모달리티 간 통합이라는 도전과제를 지닌다. 텍스트 토큰의 임베딩은 어휘 집합이라는 이산적이고 유한한 공간 내에서 정의되는 반면, 오디오 토큰 임베딩은 물리적 진동을 표현한 결과로 연속적이고 랜덤적인 특성을 지닌다. 이로 인해 오디오 임베딩은 동일한 정보에 대해서도 표현 공간 상에서 상당한 분산을 가지며, 이는 오디오 정보의 안정적인 해석에 난점이 된다.

본 논문에서 주목하는 핵심 현상은 LALM의 오디오 임베딩의 L2 노름(Norm)이 텍스트 임베딩에 비해 수백 배 크게 나타난다는 것으로, 그림 1은 MMAU 벤치마크[2]의 데이터를 통해 이를 실증한 결과이다.

본 논문에서는 이러한 노름 불균형이 LALM에서 오디오-텍스트 정렬의 준최적성(suboptimality)을 초래함을 보이고, 이를 완화하는 간단한 추론 시간 접근법인 NEAT(Norm Equalization for Audio-Text Alignment)를 제안한다. NEAT는 오디오 임베딩에 L2 노름 클리핑을 적용하여 벡터의 방향을 유지한 채 크기를 임계값 이하로 조정함으로써 두 모달리티 간 수치적 균형을 회복한다. NEAT는 Audio Flamingo 3 모델[3]에 적용했을 때 추가적인 학습 없이도 MMAU 벤치마크에서의 오디오 이해 정확도를 향상시킨다.

II. 본론

2.1 모델 및 평가 벤치마크

연구에 사용된 모델은 NVIDIA에서 개발한 오픈 소스 LALM인 Audio Flamingo 3으로, Whisper 기반 오디오 인코더인 Whisper-AF, Qwen-2.5-7B LLM 백본, 그리고 이 둘의 표현 공간을 연결하는 2층 다층 퍼셉트론으로 구성된 모달리티 프로젝터로 이루어져 있다. 평가는 음성, 환경음, 음악 이해를 포괄하는 MMAU 벤치마크의 test-mini 서브셋을 활용하였다.

2.2 NEAT

NEAT는 LALM 내부의 모달리티 프로젝터 출력 이후, LLM 입력 직전 단계에서 오디오 임베딩 벡터 e 에 대해 토큰별 L2 노름 클리핑을 적용한다. 구체적인 연산은 다음과 같이 정의된다. 여기서 τ는 클리핑 임계값, ϵ은 수치적 안정성을 위한 상수이다.

$$\hat{e} = e \cdot \frac{\min(||e||_2, \tau)}{\max(||e||_2, \varepsilon)}$$

III. 구현

본 연구에서 제안한 NEAT의 효과성 검증을 위해 클리핑 임계값을 10부터 1000까지 10 단위로 증가시키며 그리드 서치를 수행하고, 오디오 이해 벤치마크에서의 정확도를 클리핑이 없는 베이스라인과 비교하였다. 그림 2는 클리핑 임계값에 따른 MMAU-test-mini 정확도 변화를 나타내며, 두 극단 사이에 최적 임계값이 존재함을 보여준다. 표 1은 실험 조건에 따른 모달리티별 오디오 이해 정확도를 요약한 결과로, NEAT를 적용한 LALM은 각 모달리티에 대 베이스라인을 상회하였다. 특히, NEAT(Adaptive)는 각 모달리티에 독립적인 최적 임계값을 적용하는 변형으로, NEAT 대비 더 높은 성능을 달성하여 적응적 노름 조정의 효과성을 시사한다.

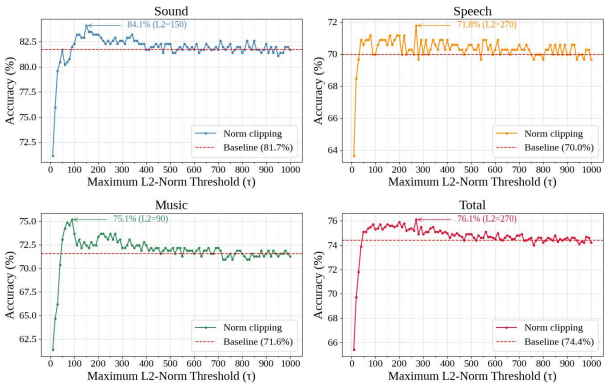

그림 2. MMAU Accuracy (%) by Threshold

	Sound	Speech	Music	Total
Baseline	81.68	69.97	71.56	74.4
NEAT	82.88	71.77	73.65	76.1
NEAT(Adaptive)	84.08	71.77	75.15	77.0

표 1. MMAU Accuracy (%) Comparison

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 LALM에서 오디오 임베딩과 텍스트 임베딩 간에 발생하는 L2 노름 불균형 문제를 제기하고, 이를 해결하기 위한 방법론인 NEAT를 제안하였다. Audio Flamingo 3 모델을 기반으로 한 실험에서 MMAU-test-mini 벤치마크에서 정확도를 향상시키는 클리핑 임계값을 그리드 서치를 통해 탐색한 결과, 적절한 노름 클리핑을 통해 LALM의 오디오 이해 정확도가 74.4%에서 76.1%로 향상할 수 있음을 확인하였다. 이 결과는 오디오-텍스트 노름 불균형이 LALM 성능의 병목 요인 중 하나임을 시사하며, 별도의 학습 없이 추론 시 간단히 적용할 수 있는 방법론으로도 LALM의 오디오 이해에 있어 의미 있는 성능 개선을 달성할 수 있음을 보여 준다.

향후 연구 방향으로는, 첫째, 고정된 단일 임계값 대신 오디오 내용의 정보량이나 의미론적 복잡도를 고려하여 토큰별로 임계값을 적응적으로 조절하는 정보량 기반 노름 조절 방법을 탐구할 것이다. 둘째, 음성의 평균 노름이 음악 및 환경음에 비해 작게 나타나는 현상을 심층 분석할 것이다. LALM의 사전 학습 데이터에서 음성이 차지하는 높은 비중을 고려할 때, 오디오 임베딩의 노름 크기가 해당 오디오 유형에 대한 모델의 학습 충분성과 역의 상관 관계를 가질 수 있다는 가설을 검증하는 것도 의미 있는 후속 연구가 될 것이다. 셋째, 다양한 LALM 및 벤치마크 상에서 NEAT의 일반화 성능 검증을 수행할 것이다.

Acknowledgement

이 논문은 2026년도 BK21 FOUR 정보기술 미래인재 교육연구단에 의해 지원되었음.

참고문헌

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in Proc. NeurIPS, 2017, pp. 5998–6008.

[2] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "MMAU: A mas300 sive multi-task audio understanding and reasoning benchmark," in Proc. ICLR, 2025.

[3] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S. gil Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro, "Audio Flamingo 3: Advancing audio intelligence with fully open large audio language models," in Proc. NeurIPS, 2025.