소형 오디오 LLM을 위한 Self-Consistency 기반 추론 성능 향상 기법에 관한 연구

홍지민, 김석민, 김남수

서울대학교 전기정보공학부 뉴미디어통신공동연구소

{jmhong, smkim}@hi.snu.ac.kr, nkim@snu.ac.kr

A Study on the Self-Consistency based reasoning performance improvement for Small Audio LLMs

Jimin Hong, Seokmin Kim, Nam Soo Kim

Department of Electrical and Computer Engineering and INMC

Seoul National University

요 약

본 논문은 소형 오디오 기반 멀티모달 언어모델의 추론 성능을 향상시키기 위해 self-consistency 기법을 적용하고, 기존 다수결 기반 투표 방식의 한계를 보완하기 위한 log probability 기반 추론 전략을 제안한다. 추가적인 학습 없이 추론 단계에서 성능을 개선하는 데 초점을 두었으며, Qwen2.5-Omni-3B 모델과 MMAU 벤치마크를 활용하여 다양한 투표 전략의 성능을 비교하였다. 실험 결과, 제안한 방법은 기존 greedy decoding 및 단순 다수결 방식 대비 추론 성능의 안정성과 정확도를 향상시키는 것을 확인하였다.

Ⅰ. 서 론

최근 멀티모달 대규모 언어모델(LLM)의 발전과 함께 텍스트뿐만 아니라 이미지, 음성, 환경음 등 오디오 정보를 이해하고 추론하는 모델의 필요성이 증가하고 있다. 이러한 오디오 기반 멀티모달 모델은 음성 비서, 스마트 디바이스, 헬스케어 등 다양한 실사용 환경에서 활용 가능성이 크다.

그러나 실제 서비스 환경이나 온디바이스(on-device) 환경에서는 연산 자원과 메모리 제약으로 인해 대규모 모델의 사용이 제한적이며, 이에 따라 소형 멀티모달 모델을 활용하면서도 최대한 높은 성능을 확보하는 것이 중요한 연구 과제로 떠오르고 있다. 특히 추가적인 학습이나 파인튜닝 없이, 추론 단계에서의 전략만으로 성능을 향상시키는 방법은 실용적인 관점에서 큰 장점을 가진다.

자연어 처리 분야에서는 이러한 목적을 위해 self-consistency 기법이 제안되어, 동일 입력에 대해 여러 추론 경로를 생성하고 이를 집계함으로써 추론 성능을 향상시키는 효과를 보였다. 그러나 오디오 기반 추론 환경에 이를 직접 적용할 경우, majority voting 과정에서 동률(tie)이 발생하는 문제로 인해 성능의 불안정성이 나타날 수 있다.

이에 본 연구에서는 기존 self-consistency 접근법을 오디오 기반 멀티모달 추론에 적용하되, 모델의 내부 log probability 정보를 활용한 voting 전략을 도입하여 동률 문제를 완화하고 보다 robust한 추론 성능 향상을 달성하고자 한다.

Ⅱ. 본론

i. Self Consistency

Self-consistency는 Chain-of-Thought(CoT) 추론 과정에서 발생할 수 있는 단일 추론 경로의 불안정성을 완화하기 위해 제안된 기법이다[1]. 기존의 greedy decoding 방식은 하나의 추론 경로만을 생성하기 때문에, 초기 추론 단계의 오류가 최종 결과에 직접적으로 영향을 미칠 수 있다. 이에 반해 self-consistency는 동일한 입력에 대해 여러 개의 추론 결과를 샘플링하고, 이들 결과를 집계하여 최종 응답을 결정함으로써 추론의 신뢰성과 정확도를 향상시킨다.

일반적으로 self-consistency는 다수결 기반 voting 방식으로 구현되며, 자연어 처리 분야에서 다양한 추론 태스크에 대해 성능 향상을 보인 바 있다. 본 연구에서는 이러한 self-consistency 개념을 오디오 기반 멀티모달 추론 환경에 적용하고, 특히 voting 과정에서 발생할 수 있는 동률 문제를 고려하여 보다 안정적인 추론 결과를 도출하는 방법을 분석한다.

ii. Audio LLM

본 연구에서는 오디오 입력을 처리할 수 있는 멀티모달 언어모델인 Qwen2.5-Omni[2] 계열 모델을 Audio LLM로 사용하였다. Qwen2.5-Omni 모델은 음성 및 다양한 오디오 신호와 텍스트 정보를 동시에 입력으로 받아 언어 기반 추론을 수행할 수 있도록 설계된 멀티모달 모델이다. 이러한 특성으로 인해 오디오 이해와 추론이 동시에 요구되는 태스크에 적합하다.

본 실험에서는 실사용 환경과 온디바이스 적용 가능성을 고려하여, 3B 규모의 소형 모델(Qwen2.5-Omni-3B)을 베이스라인으로 설정하였다. 추가적인 학습이나 파인튜닝 없이, 추론 단계에서 self-consistency 및 다양한 voting 전략을 적용함으로써 소형 Audio LLM의 추론 성능을 효과적으로 향상시킬 수 있는지를 평가하였다.

iii. Self-Consistency와 log probability를 활용한 추론 전략

일반적인 self-consistency 기법에서는 동일한 입력에 대해 여러 개의 응답을 생성한 후, 생성된 응답의 다수결을 통해 하나의 최종 답안을 도출한다. 이 과정에서 다수결 결과가 동률을 이루는 경우, 일반적으로 무작위로 최종 응답을 결정한다.

본 연구에서는 다수결 기반 self-consistency를 오디오 LLM에 적용하는 것에 더해, 이 과정에서 발생하는 동률 문제를 완화하기 위한 방법으로, log probability 정보를 활용한 추론 전략을 제안한다. 다수결 결과가 동률을 이루는 경우, 동률에 해당하는 각 선택지에 대해 생성된 응답들의 log probability 평균을 비교하고, 가장 높은 값을 갖는 선택지를 최종 응답으로 선택한다.

제안한 방법은 단순히 생성된 응답의 빈도만을 고려하는 것이 아니라, 각 응답이 모델에 의해 얼마나 신뢰도 높게 생성되었는지를 함께 반영한다는 점에서 의미를 가진다. 이를 통해 동률 상황에서 보다 합리적인 응답 선택이 가능하며, self-consistency 기반 추론의 안정성과 성능 향상을 기대할 수 있다.

iv. 평가 방법

본 연구에서는 추론 성능 평가를 위해 MMAU (Massive Multi-task Audio Understanding)[3] 벤치마크 중 MMAU-test-mini 데이터셋을 사용하였다. MMAU는 speech, sound, music 등 다양한 오디오 입력을 기반으로 한 문제들로 구성되어 있으며, 총 1,000개의 4지선다형(multiple-choice) 데이터셋을 제공한다. 본 실험에서는 해당 벤치마크를 활용하여 self-consistency 기법의 효과를 Information Extraction과 Reasoning 카테고리의 accuracy(%) 기준으로 분석하였다.

제안한 방법의 효과를 검증하기 위해, 비교 기준으로 deterministic한 greedy decoding 기반 추론 방식을 사용하였다. 또한 기본적인 self-consistency 기법으로, 다수결 기반 voting만을 적용한 방식과의 성능을 비교하였다.

추가적으로, 다수결을 사용하지 않고 생성된 모든 응답에 대해 선택지별 평균 log probability를 계산하여 최종 응답을 결정하는 방식과도 비교 실험을 수행하였다.

v. 실험 결과

기본적으로 Self-Consistency를 적용했을 때, greedy decoding에 비해 성능이 올라가는 것이 확인되어 audio LLM에 이 방식이 효과적이라는 것을 보여준다. 이에 더해 다수결 과정에서 동률이 발생했을 때 무작위로 응답을 선택하는 방식보다 log probability를 이용하는 것이 선택의 신뢰성이 반영되어 성능이 추가적으로 향상되는 것으로 분석된다. 다만 N=10 설정에서는 생성 응답 수가 증가함에 따라 다수결 결과가 상대적으로 안정화되어, 제안한 방식의 효과가 제한적으로 나타난다.

한편, 다수결을 사용하지 않고 log probability만을 기준으로 응답을 선택한 경우, 전반적으로 성능 저하가 발생하는 것을 확인하였다. 이는 self-consistency에서 다수결이 1차적인 필터 역할을 수행하며, 단순 확률 정보만으로는 추론의 안정성을 확보하기 어렵다는 점을 시사한다. 이를 통해 다수결을 기반으로 하되 log probability를 보조적으로 활용하는 방식이 상호보완을 통한 가장 효과적인 전략임을 확인하였다.

		Information Extraction	Reasoning
greedy		63.78	65.14
N=5	Majority + random(tie)	66.87	63.81
	Average LogProb	64.09	60.41
	Majority + logprob(tie) **(ours)**	**67.18**	**65.88**
N=10	Majority + random(tie)	68.73	**66.03**
	Average LogProb	60.37	59.53
	Majority + log prob(tie) **(ours)**	**69.04**	65.88

표 1 각 추론 전략에 따른 MMAU 벤치마크에 대한 성능 비교

Ⅲ. 결론

기본적으로 Self-Consistency를 적용했을 때, greedy decoding에 비해 성능이 올라가는 것이 확인되어 audio LLM에 이 방식이 효과적이라는 것을 보여준다. 이에 더해 다수결 과정에서 동률이 발생했을 때 무작위로 응답을 선택하는 방식보다 log probability를 이용하는 것이 선택의 신뢰성이 반영되어 성능이 추가적으로 향상되는 것으로 분석된다. 다만 N=10 설정에서는 생성 응답 수가 증가함에 따라 다수결 결과가 상대적으로 안정화되어, 제안한 방식의 효과가 제한적으로 나타난다.

한편, 다수결을 사용하지 않고 log probability만을 기준으로 응답을 선택한 경우, 전반적으로 성능 저하가 발생하는 것을 확인하였다. 이는 self-consistency에서 다수결이 1차적인 필터 역할을 수행하며, 단순 확률 정보만으로는 추론의 안정성을 확보하기 어렵다는 점을 시사한다. 이를 통해 다수결을 기반으로 하되 log probability를 보조적으로 활용하는 방식이 상호보완을 통한 가장 효과적인 전략임을 확인하였다.

ACKNOWLEDGMENT

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

참 고 문 헌

[1] X. Wang et al., “Self-Consistency Improves Chain of Thought Reasoning in Language Models.” 2023. [Online]. Available: https://arxiv.org/abs/2203.11171

[2] Jin Xu, “Qwen2.5-Omni Technical Report,” arXiv preprint arXiv:2503.20215, 2025.

[3] S. Sakshi et al., “MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark.” 2024. [Online]. Available: https://arxiv.org/abs/2410.19168