누락된 모달리티에 강건한 멀티모달 분류를 위한 Prompt Learning 기반 모델 학습

김준섭(성균관대학교) | 박은주(이화여자대학교) | 양희재(성균관대학교)

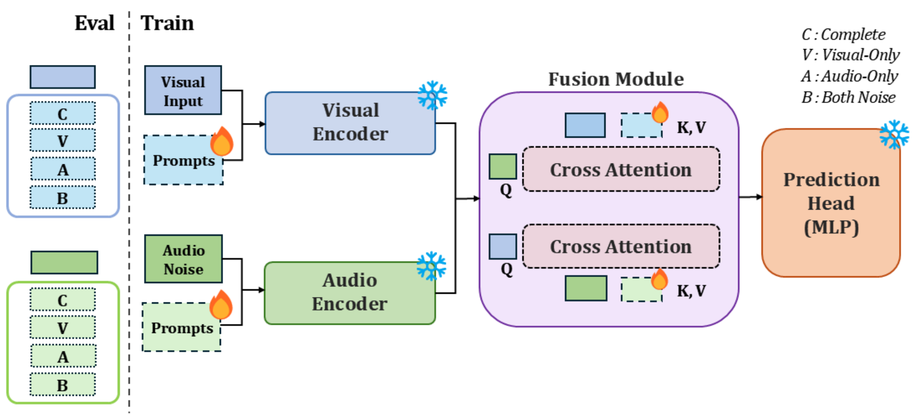
ABSTRACT

Multimodal classification systems often suffer from severe performance drops when one or more input modalities are missing or corrupted. To address this issue, we propose a lightweight and robust framework based on prompt learning. Our method introduces learnable prompt tokens at both the input and attention levels, enabling the model to capture reliable cross-modal representations even in challenging conditions.

By simulating various degradation scenarios during training, the model gains resilience to missing or noisy inputs. Experiments on the UrbanSound8K-AV and CIFAR10-AV datasets demonstrate that our approach outperforms existing methods—such as CAV-MAE, LoRA, and Adapter—achieving up to 10.4% higher accuracy, while reducing training time by 96% and memory usage by 82.3%. These results highlight the practicality and efficiency of prompt learning for robust multimodal classification in real-world settings.

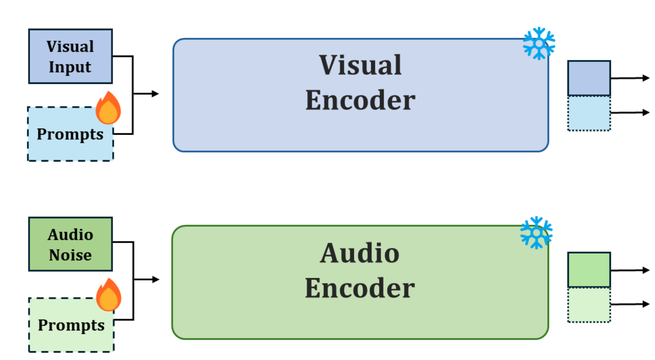
METHOD

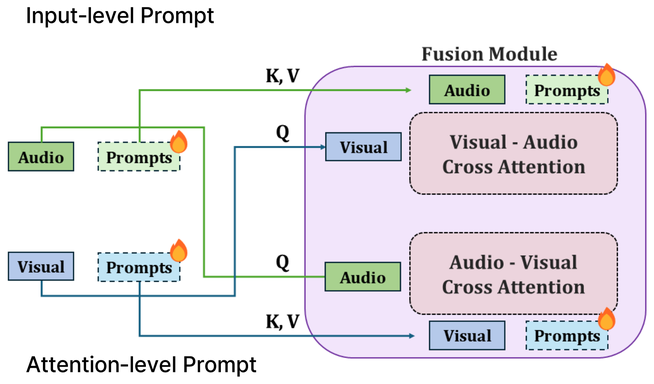
Framework



본연구는 결손된 모달리티 환경에서도 강건한 멀티모달 분류를 위해 Prompt Learning 기반 프레임워크 제안

- **입력 및 어텐션 단계에 프롬프트 삽입**
  학습 가능한 프롬프트 토큰을 입력과 어텐션 Key/Value에 삽입하여, 손상된 모달리티에서도 일관된 표현을 유도
- **Cross-Attention 기반 양방향 Fusion 구조**
  오디오-비주얼 간 상호 보완적 정보를 교환하는 양방향 Cross-Attention 구조를 통해 표현 정렬을 강화
- **손상 시나리오별 프롬프트 학습 전략**
  정상, 오디오 손상, 비주얼 손상, 양쪽 손상 등 다양한 조건별 프롬프트를 독립적으로 학습하여, 테스트 시 조건 없이도 강건한 추론이 가능하도록 설계





Prompt Mechanism

- Input-Level Prompt

오디오 및 비주얼 입력 각각에 대해, 프롬프트 토큰을 앞단에 삽입하여 각 모달리티의 신뢰도와 중요도를 반영한 표현을 학습

> 비주얼 입력이 손상된 경우에도 오디오 프롬프트를 통해 보완적 정보를 활용할 수 있도록 설계

- Attention-Level Prompt

프롬프트 토큰은 단순 입력 보조 역할을 넘어, Transformer의 어텐션 연산 내 Key / Value에 통합되어 모달리티 간 상호작용을 동적으로 조절.

> 즉, 어텐션이 어떤 모달리티에 집중할지를 프롬프트가 동적으로 조정하며, 손상된 입력에서는 보다 신뢰도 높은 모달리티로 집중 유도할 수 있도록 설계

- Cross-Attention Fusion Module

본 모듈은 양방향 Cross-Attention 구조를 기반으로, 시각 및 오디오 특성 간의 정렬을 학습

- Visual-to-Audio: 비주얼 쿼리가 오디오 + 오디오 프롬프트에 주의
- Audio-to-Visual: 오디오 쿼리가 비주얼 + 비주얼 프롬프트에 주의

이를 통해 하나의 모달리티가 손상되어도 다른 모달리티가 보완할 수 있으며, 전체적인 표현 안정성을 유지.

RESULT

Comparison Accuracy

> Urbansound8K-AV

Noise Type	CAV-MAE	LoRA	Adapter	Prompt Learning (Ours)
Complete (C)	0.99	0.98	0.98	**0.99**
Noise to Audio (A)	0.69	0.84	0.88	**0.94**
Noise to Vision (V)	0.83	**0.84**	0.81	0.83
Noise to Both (D)	0.71	0.80	0.81	**0.88**
Avg.	0.80	0.87	0.87	**0.91**

> Cifar10-AV

Noise Type	CAV-MAE	LoRA	Adapter	Prompt Learning (Ours)
Complete (C)	0.93	0.92	0.91	**0.93**
Noise to Audio (A)	0.66	0.77	0.81	**0.89**
Noise to Vision (V)	0.70	0.76	**0.85**	0.84
Noise to Both (D)	0.60	**0.79**	0.72	0.78
Avg.	0.72	0.81	0.82	**0.86**

- 본 실험에서는 다양한 모달리티 손상 조건(C, A, V, D)에서 모델의 강건성을 비교
- UrbanSound8K-AV에서 0.91, CIFAR10-AV에서 0.86의 평균 정확도 기록(최고성능)
- 특히 오디오 손상(A) 및 양쪽 손상 조건(D)에서 기존 방법 대비 큰 향상을 보여, 프롬프트 기반 학습이 결손 대응 및 표현 정합성 유지에 효과적임을 입증

Efficient Analysis

Methods	Trainable Params (M)	Total Memory (GiB)	Time / Epoch	Accuracy
Finetuning	86.4	95.1	60.0 sec	0.88
LoRA	12.6	27.2	10.1 sec	0.86
Adapter	15.4	32.8	12.5 sec	0.87
Prompt Learning (Ours)	**3.2**	**17.8**	**2.4 sec**	**0.88**

- 본 실험은 Prompt Learning이 기존 Fine-Tuning, LoRA, Adapter 방식 대비 성능과 자원 효율성 모두에서 우수함을 입증하기 위해 수행
- 제안된 방식은 전체 파라미터의 3.7% 수준(3.2M)만 학습하고도, 정확도 0.88로 Fine-Tuning과 동일한 성능을 달성
- 또한, 본 방식은 학습 시간을 96% 감소 (60.0s → 2.4s) , 메모리 사용량을 82.3% 절감 (95.1GiB → 17.8GiB)
- 이와 같은 성능 절감에도 불구하고, 백본을 고정한 상태에서 정확도를 유지하며 고효율 학습 구조로서의 가능성을 입증
- 이는 제한된 자원 환경에서도 적용 가능한 실용적이고 확장 가능한 접근법으로,향후 다양한 멀티모달 태스크에 경량화된 대안으로 활용될 수 있음을 시사

CONCLUSION

- 본 연구는 일부 모달리티가 결손되거나 손상된 환경에서도 강건한 분류 성능을 보이는 멀티모달 프레임워크를 제안
- 입력 및 어텐션 단계의 프롬프트 삽입, Cross-Attention 기반 Fusion Module, 그리고 시나리오별 프롬프트 학습 전략을 결합하여 UrbanSound8K-AV와 CIFAR10-AV에서 높은 정확도와 효율성을 동시에 입증
- 그러나 본 모델은 주요 오디오-비주얼 벤치마크에 한정되어 검증되었으며, 보다 다양한 도메인과 모달리티 조건에서의 일반화 성능은 추가 연구가 필요
- 향후에는 프롬프트 구조의 확장 및 다중 모달리티가 복합적으로 손상되는 시나리오에 대한 강건성 향상을 중심으로 아키텍처를 발전시킬 예정

REFERENCE

[1] Lee, Yi-Lun, et al. "Multimodal prompting with missing modalities for visual recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

[2] Gong, Yuan, et al. "Contrastive audio-visual masked autoencoder." arXiv preprint arXiv:2210.07839, 2022.

[3] Guo, Lingyue, et al. "Transformer-based spiking neural networks for multimodal audiovisual classification." IEEE Transactions on Cognitive and Developmental Systems 16.3 (2023): 1077-1086.

[4] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." Proceedings of the 22nd ACM international conference on Multimedia. 2014.