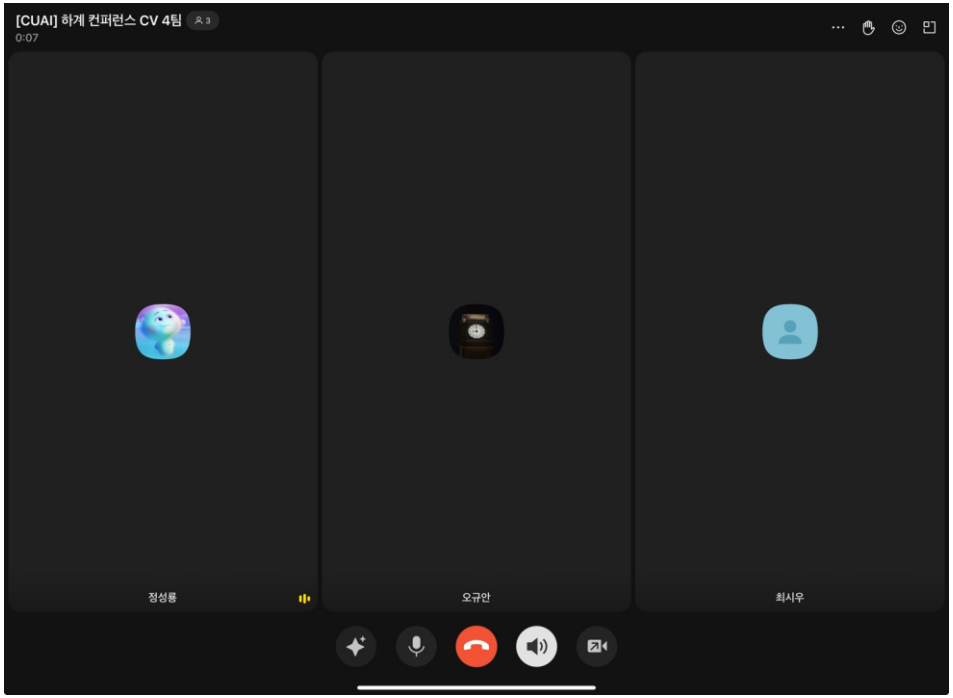
CUAI Multimodal 논문 리뷰 스터디팀

2024.11.12

발표자 : 오규안

스터디원 소개 및 만남 인증



스터디원 1 : 오규안 (AI)

스터디원 2 : 최시우 (AI)

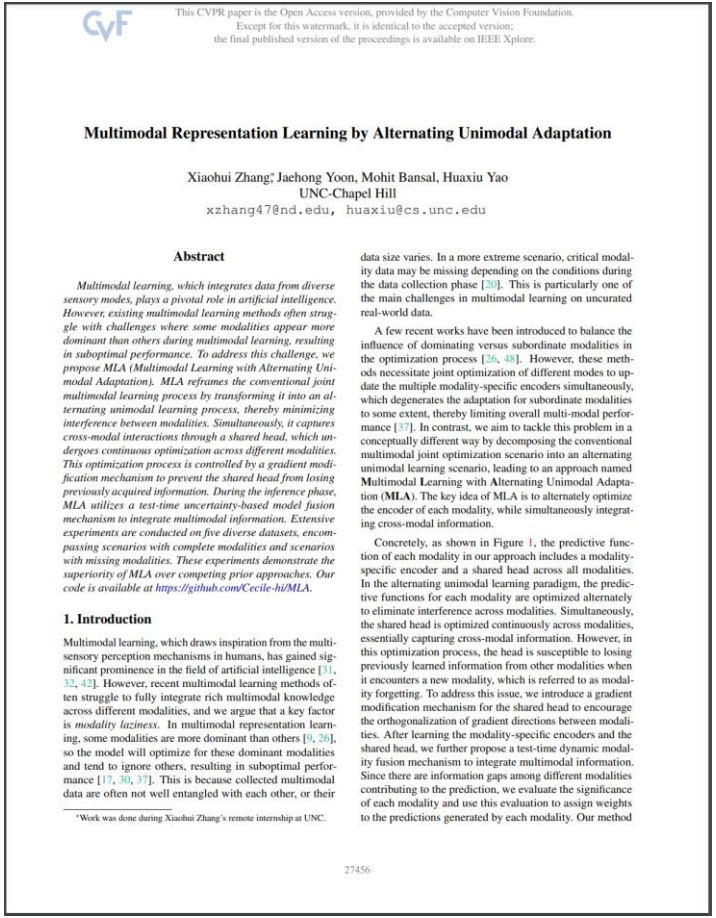
스터디원 3 : 정성룡 (AI)

논문 선정

“Multimodal Representation Learning by Alternating Unimodal Adapation”

(CVPR 2024)

“단일 모달 적응 형식의 다중 모달 표현 학습“



This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multimodal Representation Learning by Alternating Unimodal Adaptation

Xiaohui Zhang* Jaehong Yoon, Mohit Bansal, Huaxiu Yao
UNC-Chapel Hill
xzhang47@nd.edu, huaxiu@cs.unc.edu

Abstract

Multimodal learning, which integrates data from diverse sensory modes, plays a pivotal role in artificial intelligence. However, existing multimodal learning methods often struggle with challenges where some modalities appear more dominant than others during multimodal learning, resulting in suboptimal performance. To address this challenge, we propose MLA (Multimodal Learning with Alternating Unimodal Adaptation). MLA reframes the conventional joint multimodal learning process by transforming it into an alternating unimodal learning process, thereby minimizing interference between modalities. Simultaneously, it captures cross-modal interactions through a shared head, which undergoes continuous optimization across different modalities. This optimization process is controlled by a gradient modification mechanism to prevent the shared head from losing previously acquired information. During the inference phase, MLA utilizes a test-time uncertainty-based model fusion mechanism to integrate multimodal information. Extensive experiments are conducted on five diverse datasets, encompassing scenarios with complete modalities and scenarios with missing modalities. These experiments demonstrate the superiority of MLA over competing prior approaches. Our code is available at https://github.com/Cecile-hi/MLA.

1. Introduction

Multimodal learning, which draws inspiration from the multisensory perception mechanisms in humans, has gained significant prominence in the field of artificial intelligence [31, 32, 42]. However, recent multimodal learning methods often struggle to fully integrate rich multimodal knowledge across different modalities, and we argue that a key factor is *modality laziness*. In multimodal representation learning, some modalities are more dominant than others [9, 26], so the model will optimize for these dominant modalities and tend to ignore others, resulting in suboptimal performance [17, 30, 37]. This is because collected multimodal data are often not well entangled with each other, or their data size varies. In a more extreme scenario, critical modality data may be missing depending on the conditions during the data collection phase [20]. This is particularly one of the main challenges in multimodal learning on uncurated real-world data.

A few recent works have been introduced to balance the influence of dominating versus subordinate modalities in the optimization process [26, 48]. However, these methods necessitate joint optimization of different modes to update the multiple modality-specific encoders simultaneously, which degenerates the adaptation for subordinate modalities to some extent, thereby limiting overall multi-modal performance [37]. In contrast, we aim to tackle this problem in a conceptually different way by decomposing the conventional multimodal joint optimization scenario into an alternating unimodal learning scenario, leading to an approach named **M**ultimodal **L**earning with **A**lternating Unimodal Adaptation (**MLA**). The key idea of MLA is to alternately optimize the encoder of each modality, while simultaneously integrating cross-modal information.

Concretely, as shown in Figure 1, the predictive function of each modality in our approach includes a modality-specific encoder and a shared head across all modalities. In the alternating unimodal learning paradigm, the predictive functions for each modality are optimized alternately to eliminate interference across modalities. Simultaneously, the shared head is optimized continuously across modalities, essentially capturing cross-modal information. However, in this optimization process, the head is susceptible to losing previously learned information from other modalities when it encounters a new modality, which is referred to as modality forgetting. To address this issue, we introduce a gradient modification mechanism for the shared head to encourage the orthogonalization of gradient directions between modalities. After learning the modality-specific encoders and the shared head, we further propose a test-time dynamic modality fusion mechanism to integrate multimodal information. Since there are information gaps among different modalities contributing to the prediction, we evaluate the significance of each modality and use this evaluation to assign weights to the predictions generated by each modality. Our method

*Work was done during Xiaohui Zhang's remote internship at UNC.

01. Background & Problems

I. **Multimodal Learning:** 다양한 유형의 데이터를 (텍스트, 이미지, 오디오, 텍스트 등) 동시에 처리하고, 그들 간의 관계를 학습하여 정보 표현을 추구하는 접근법

II. **Problem: Modality Laziness 문제**

- 최근의 Multimodal Learning 방법들은 여러 Modalities의 정보를 효과적으로 통합하는데 많은 어려움을 겪고 있습니다.
- 특정 모드가 더 우세하게 작용하면, 그 모델이 그 모드에 최적화되고, 다른 모드는 무시되는 경향이 생깁니다.
- Ex) 이미지 & 텍스트를 동시에 다룰 때, 이미지 모드가 텍스트보다 성능에 더 큰 영향을 미쳐, 텍스트 정보를 제대로 활용하지 못하는 경우

02. Limitations of past approaches

I. **기존의 Multimodal Learning 방식들은 모드 간의 영향 균형을 맞추기 위해,** 다양한 방법들을 제시하였지만, 대부분 여러 모드를 동시에 최적화해야 하므로, **부차적인 모드 (dominant 모드가 아닌 모드) 에 대한 적응이 부족해지는 상황이** 발생합니다.

II. **공동 최적화 방식:** 모든 모드에 대해서 하나의 모델을 동시에 학습시키는데, **부차적인 모드를 충분히 고려하지 않아서 성능이 제한적**입니다.

III. 본 논문에서 제안하는 해결책:

"Multimodal Learning with Alternating Unimodal Adaptation"

"MLA"를 제안합니다.

03. Main ideas of MLA

I. **MLA (Multimodal Learning with Alternating Unimodal Adaptation)**

- 기존의 다중 모드 공동 최적화 방식을 **Alternating Unimodal adaptation (교차 단일 모드 학습) 으로 분해**하여, **각 모드를 독립적으로 최적화**하면서도, **모드 간 상호작용을 보존하는 방법을 제시**합니다.

II. **각 모드에 대해 최적화된 Encoder를 교차적으로 학습시키면서도, 각 모드에서 공유되는 Shared head를 사용하여 모드 간 정보를 통합합니다.**

- 이를 통해, 각 모드의 독립적인 학습을 보장하면서도, **모든 모드 간의 교차적인 정보 통합을 가능**하게 해줍니다.

04. Structure of MLA

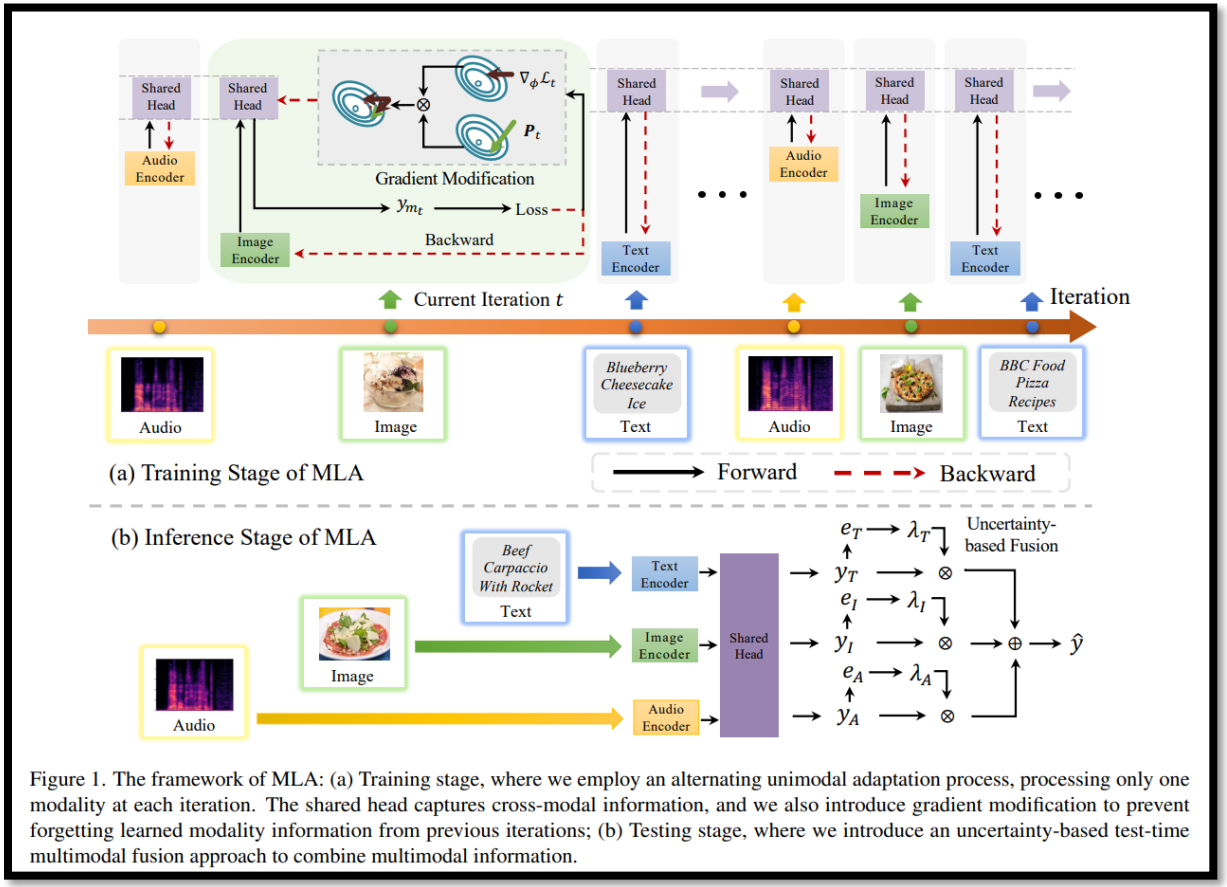


Figure 1. The framework of MLA: (a) Training stage, where we employ an alternating unimodal adaptation process, processing only one modality at each iteration. The shared head captures cross-modal information, and we also introduce gradient modification to prevent forgetting learned modality information from previous iterations; (b) Testing stage, where we introduce an uncertainty-based test-time multimodal fusion approach to combine multimodal information.

05. Alternating Unimodal adaptation

I. 교차 단일 모드의 장점

i. 모드 간의 간섭을 제거: 각 모드의 예측 함수는 번갈아가며 최적화되고, 한 모드가 다른 모드의 학습을 방해하는 일이 줄어듭니다.

ii. 교차 모드 정보 통합: 공유 헤드는 모든 모드에 대해 지속적으로 최적화되며, 이 과정에서 다중 모드 간 정보를 통합하여, 보다 정교한 예측을 가능하게 합니다.

II. 교차 단일 모드의 단점

- 교차 모드 정보를 통합하는 과정에서 **Modality Forgetting (모드 망각)**이 발생할 수 있습니다.

새로운 모드를 만나면, 이전에 학습한 다른 모드의 정보가 잊혀지는 문제가 발생할 수 있습니다.

06. Gradient Modification Mechanism

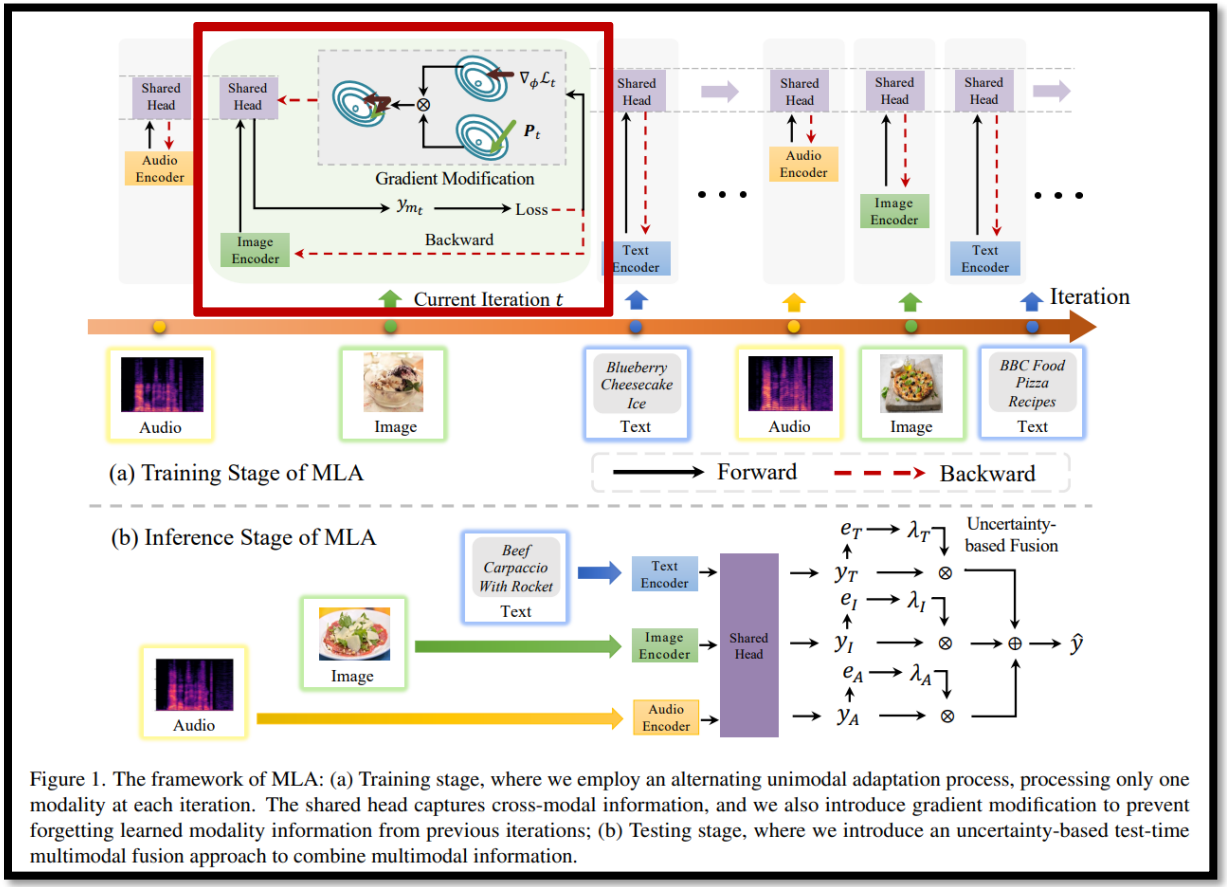


Figure 1. The framework of MLA: (a) Training stage, where we employ an alternating unimodal adaptation process, processing only one modality at each iteration. The shared head captures cross-modal information, and we also introduce gradient modification to prevent forgetting learned modality information from previous iterations; (b) Testing stage, where we introduce an uncertainty-based test-time multimodal fusion approach to combine multimodal information.

06. Gradient Modification Mechanism

I. Gradient 수정: 이 문제를 해결하기 위해, Gradient Modification Mechanism을 도입하여, 각 모드 간의 Gradient 방향을 **Orthogonalization (직교화) 시킵니다.**

II. **모드 간 간섭을 줄이고, 모드 간 정보가 상호 배타적이지 않도록 보장합니다.**

Process) 이미지와 텍스트 인코더의 학습 과정에서, 이미지 모드와 텍스트 모드의 Gradient들은 서로 직교화해야 합니다. Inner Product를 0으로 유도하여, 가중치 갱신 시 각 모드에 대한 Gradient 방향을 조정하거나, 재구성하는 방식을 적용합니다. 이를 통해, 텍스트 인코더가 이미지 인코더의 Gradient를 방해하지 않도록 하며, 한 모드의 학습이 다른 모드의 정보를 잃어버리지 않도록 합니다. **(Modality Forgetting 문제 해결)**

07. Test-Time Dynamic Modality Fusion

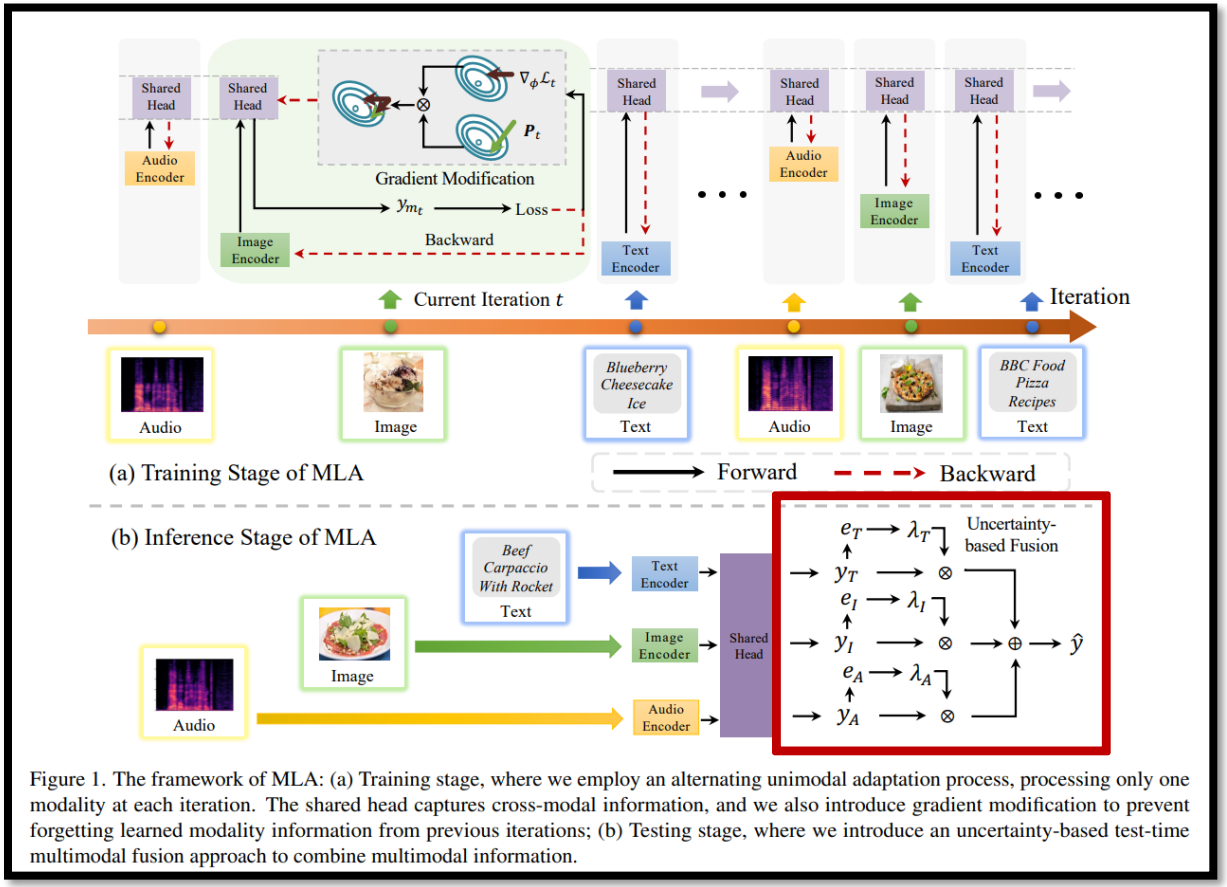


Figure 1. The framework of MLA: (a) Training stage, where we employ an alternating unimodal adaptation process, processing only one modality at each iteration. The shared head captures cross-modal information, and we also introduce gradient modification to prevent forgetting learned modality information from previous iterations; (b) Testing stage, where we introduce an uncertainty-based test-time multimodal fusion approach to combine multimodal information.

07. Test-Time Dynamic Modality Fusion

I. Test 시간에 동적으로 모드 정보를 융합하는 방법입니다. 이 방법은 각 모달리티에서 나온 예측의 신뢰도를 기반으로 가중치를 동적으로 조정하여, **불확실성이 큰 모드에 대해서는 가중치를 낮추고, 불확실성이 적은 모드의 예측에 대해서는 더 높은 가중치를 부여**합니다. 이를 통해, 잘못된 예측을 최소화하고, 예측의 정확성을 향상시킬 수 있습니다.

$$\hat{y}_r = \sum_{m=1}^{M} \lambda_{m,r} f_m(x_{m,r}; \theta_m^*, \phi^*),$$

$$e_{m,r} = -p_{m,r}^T \log p_{m,r},$$
$$\text{where } p_{m,r} = \text{Softmax}(f_m(x_{m,r}; \theta_m^*, \phi^*)).$$

$$\lambda_{m,r} = \frac{\exp\left(\max_{m=1,\ldots,M} e_{m,r} - e_{m,r}\right)}{\sum_{v=1}^{M} \exp\left(\max_{m=1,\ldots,M} e_{m,r} - e_{v,r}\right)}.$$

08. Experiment

Table 1. Results on audio-video (A-V), image-text (I-T), and audio-image-text (A-I-T) datasets. Both the results of only using a single modality and the results of combining all modalities ("Multi") are listed. We report the average test accuracy (%) of three random seeds. Full results with standard deviation are reported in Appendix A.4. The best results and second best results are **bold** and underlined, respectively.

Type	Data		Sum	Concat	Late Fusion	FiLM	BiGated	OGM-GE	QMF	**MLA (Ours)**
A-V	CREMA-D	Audio	54.14	55.65	52.17	53.89	51.49	53.76	**59.41**	59.27
		Video	18.45	18.68	55.48	18.67	17.34	28.09	39.11	**64.91**
		Multi	60.32	61.56	66.32	60.07	59.21	68.14	63.71	**79.70**
	KS	Audio	48.77	49.18	47.87	48.67	49.96	48.87	51.57	**54.67**
		Video	24.53	24.67	46.76	23.15	23.77	29.73	32.19	**51.03**
		Multi	64.72	64.84	65.53	63.33	63.72	65.74	65.78	**71.35**
I-T	Food-101	Image	4.57	3.51	58.46	4.68	14.20	22.35	45.74	**69.60**
		Text	85.63	86.02	85.19	85.84	85.79	85.17	84.13	**86.47**
		Multi	86.19	86.32	90.21	87.21	88.87	87.54	92.87	**93.33**
	MVSA	Text	73.33	75.22	72.15	74.85	73.13	74.76	74.87	**75.72**
		Image	28.46	27.32	45.24	27.12	28.15	31.98	32.99	**54.99**
		Multi	76.19	76.25	76.88	75.34	75.94	76.37	77.96	**79.94**
A-I-T	IEMOCAP	Audio	39.79	41.93	43.12	41.64	42.23	41.38	42.98	**46.29**
		Image	29.44	30.00	32.38	29.85	27.45	30.24	31.22	**37.63**
		Text	65.16	67.84	68.79	66.37	65.16	70.79	75.03	**73.22**
		Multi	74.18	75.91	74.96	74.32	73.34	76.17	76.17	**78.92**

08. Experiment

Table 2. We report the test accuracy percentages (%) on the IEMOCAP dataset using three different seeds, while applying varying modality missing rates to audio, image, and text data. The best results are highlighted in **bold**, while the second-best results are underlined.

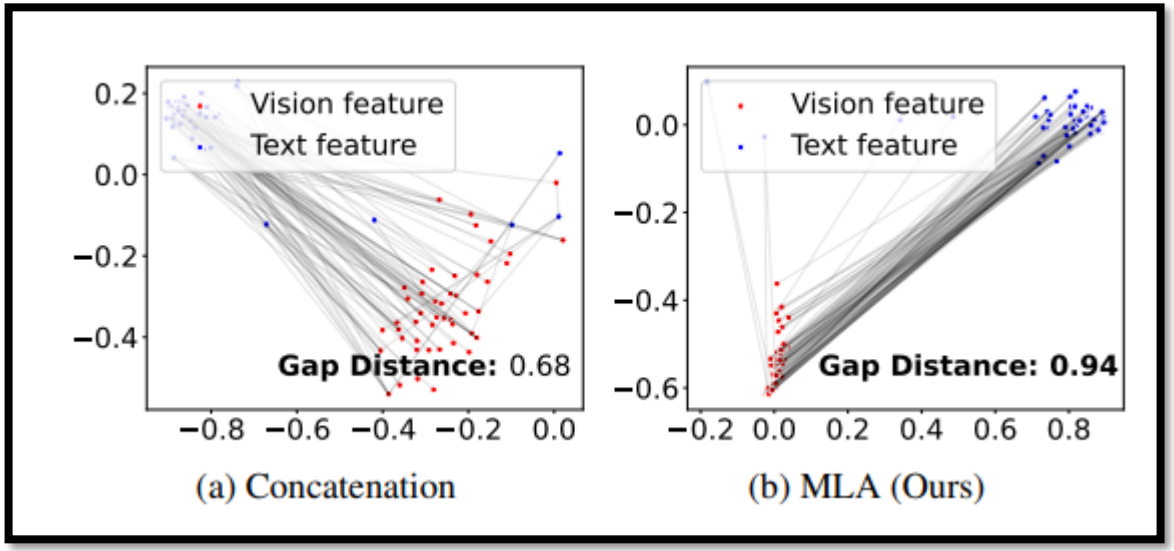
Method	Modality Missing Rate (%)						
	10	20	30	40	50	60	70
Late Fusion	72.95	69.06	64.89	61.09	56.48	52.41	45.07
QMF	73.49	71.33	65.89	62.27	57.94	55.60	50.25
CCA	65.19	62.60	59.35	55.25	51.38	45.73	30.61
DCCA	57.25	51.74	42.53	36.54	34.82	33.65	41.09
DCCAE	61.66	57.67	54.95	51.08	45.71	39.07	41.42
AE	71.36	67.40	62.02	57.24	50.56	43.04	39.86
CRA	71.28	67.34	62.24	57.04	49.86	43.22	38.56
MMIN	71.84	69.36	66.34	63.30	60.54	57.52	55.44
IF-MMIN	71.32	68.29	64.17	60.13	57.45	53.26	52.04
CPM-Net	55.29	53.65	52.52	51.01	49.09	47.38	44.76
TATE	67.84	63.22	62.19	60.36	58.74	57.99	54.35
MLA (Ours)	**75.07**	**72.33**	**68.47**	**67.00**	**63.48**	**59.17**	**55.89**

08. Experiment

Table 3. Results of ablation studies on five datasets. We report the average test accuracy (%) of three random seeds. Full results with standard deviation are reported in Appendix A.4. Note that dynamic fusion is only applied in the multimodal fusion process, which does not affect the performance of using a single modality. HGM: head gradient modification; DF: dynamic fusion.

Data		HGM	DF	HGM	DF	HGM	DF	HGM	DF
		✗	✗	✗	✓	✓	✗	✓	✓
CREMA-D	Audio	52.17		52.17		**59.27**		**59.27**	
	Video	55.48		55.48		**64.91**		**64.91**	
	Multi	66.32		72.79		74.51		**79.70**	
KS	Audio	47.87		47.87		**54.66**		**54.66**	
	Video	46.76		46.76		**51.03**		**51.03**	
	Multi	65.53		66.34		70.72		**71.35**	
Food-101	Text	85.19		85.19		**86.47**		**86.47**	
	Image	58.46		58.46		**69.60**		**69.60**	
	Multi	90.21		91.37		91.72		**93.33**	
MVSA	Text	72.15		72.15		**75.72**		**75.72**	
	Image	45.24		45.24		**54.99**		**54.99**	
	Multi	76.88		77.53		79.59		**79.94**	
IEMOCAP	Audio	43.12		43.12		**46.29**		**46.29**	
	Text	68.79		68.79		**73.22**		**73.22**	
	Image	32.38		32.38		**37.63**		**37.63**	
	Multi	74.96		75.42		77.58		**78.92**	

08. Experiment



09. Analysis

I. 문제 해결: MLA는 다중 모드 학습의 주요 문제였던 **모드 게으름을 해결**하고, **독립적인 교차 단일 모드 학습**을 통해 **모든 모드 간의 균형을 유지하며 성능을 향상**시켰습니다.

II. 주요 기여: 각 모드의 독립적인 최적화와 공유 헤드를 통한 정보 통합을 통해, **MLA는 모드 간의 간섭을 줄이면서도 모드 망각 문제를 해결**했습니다.

III. 동적 모드 융합: 시험 시 각 모드의 신뢰도를 평가해 **가중치를 동적으로 조정하여**, 불확실성이 큰 모드의 영향을 줄이고 예측의 정확성을 높였습니다.

IV. 실험 결과: MLA는 기존의 방법들과 비교해 더욱 뛰어난 성능을 보였으며, 특히 모드 누락 상황에서도 강력한 성능을 유지함을 입증했습니다.

V. 추가 개선 여지: 다만, 모드 망각 문제는 완벽하게 해결되지 않은 것으로 보입니다. 이 부분은 향후 연구에서 좀 더 발전시킬 수 있는 분야이며, **Gradient Modification Mechanism의 개선이나 추가적인 정교화가** 필요하다고 생각합니다.

감사합니다