Bridging the Gap between Heterogeneous Data: Multimodal Representation Alignment via Disentangled Learning

2025년 9월 5일

박현우

발표자 소개

❖ 박현우 (Hyunwoo Park)

- 고려대학교 산업경영공학과 석사과정 (2025.03 ~ Present)
- Data Mining & Quality Analytics Labs. (김성범 교수님)

❖ Research Interest

- Multimodal Learning

❖ Contact

- phwnob20@korea.ac.kr

Seminar Outline

- Introduction
  - Multimodal learning

- Disentangled Representation Learning Methods for Multimodal Data
  - Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)
  - Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)
  - Triple Disentangled Representation Learning for Multimodal Affective Analysis (2024, Information Fusion)

- Summary

Introduction

What is multi-modal learning and why we need it

❖ Multi-modal learning

- 여러 종류의 모달리티를 함께 사용해 모델을 학습하는 방법
- 텍스트, 이미지, 음성, 비디오 등이 각각의 모달리티가 될 수 있음

Introduction

What is multi-modal learning and why we need it

❖ Why multi-modal for AI?

- 현실 세계의 많은 문제들은 하나의 모달리티로는 완벽하게 설명하기 어려움
- 감정 분석 문제(Sentiment Analysis)

Unfortunately, it’s also the new worst product I think I’ve ever reviewed in its current state.

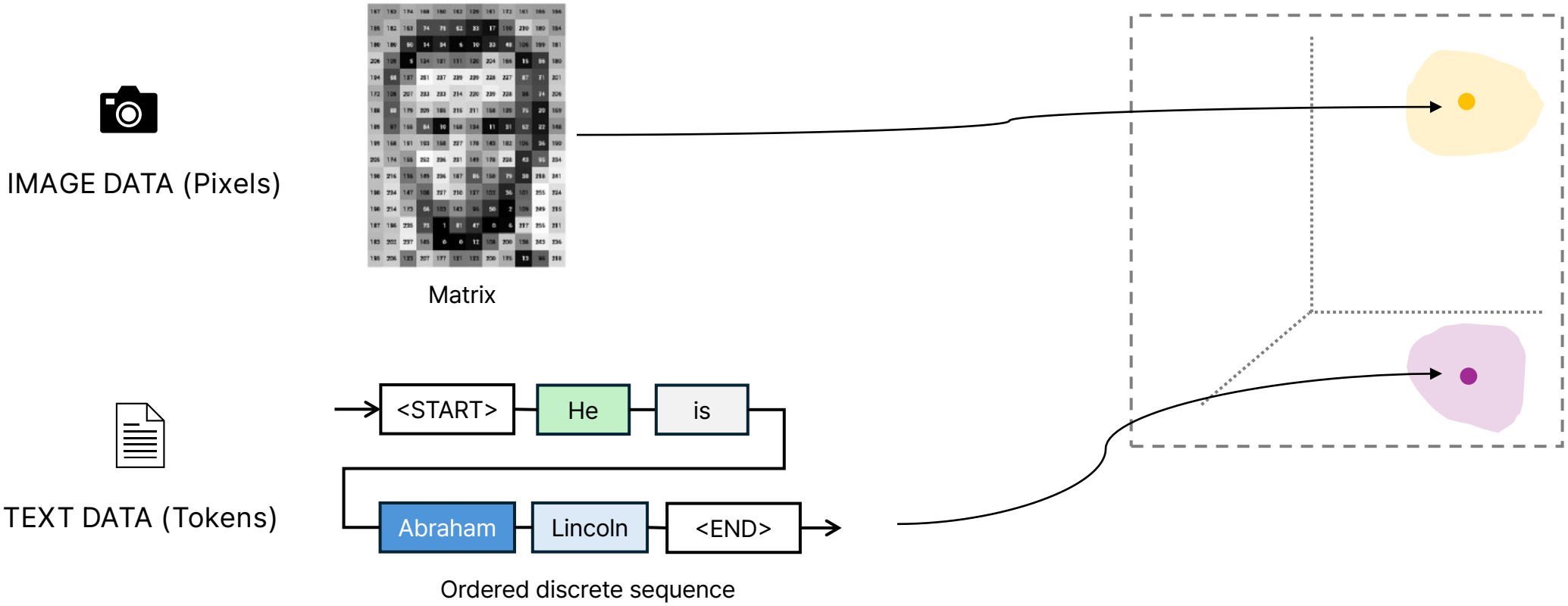
Marques Brownlee - YouTube

Introduction

The heterogeneity gap

❖ 이질성 문제

- 각 모달리티 데이터는 근본적으로 다른 구조와 통계적 속성을 가짐
- 이질성을 고려하지 않고 활용하게 될 경우 성능 저하



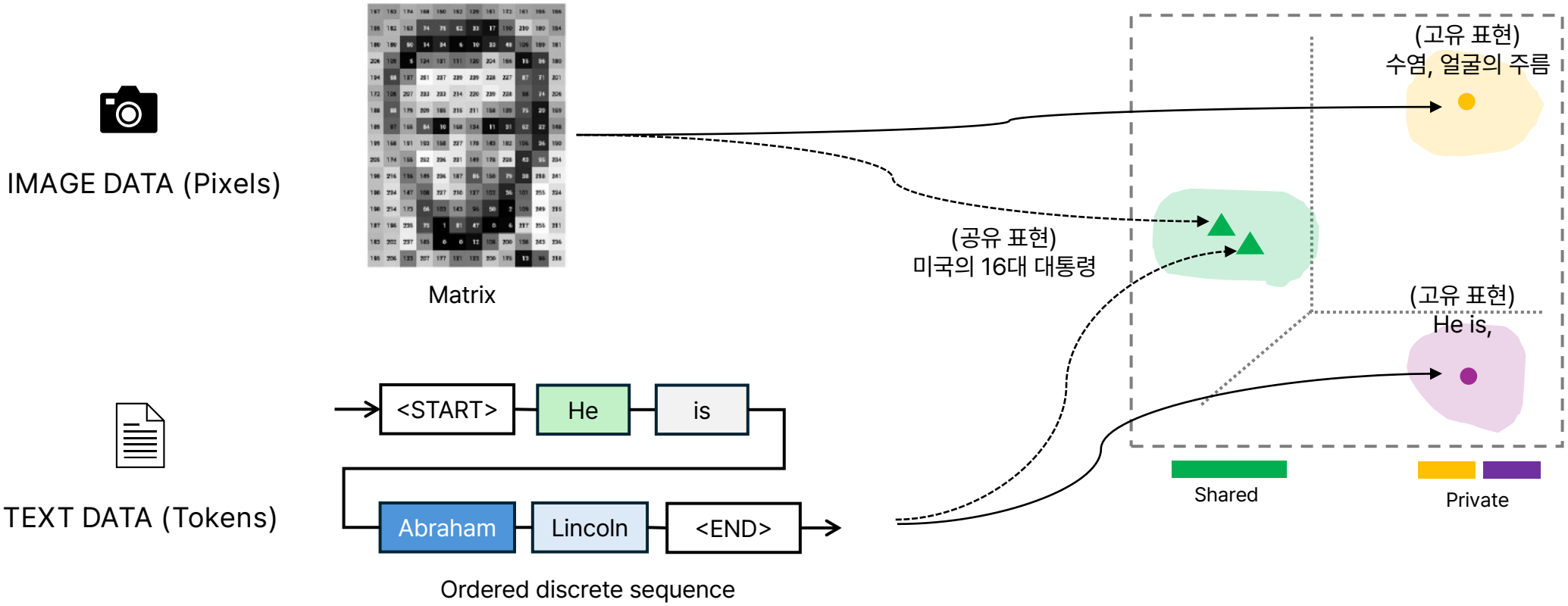
-Pixel Data Diagram of Abraham Lincoln 166 | Download Scientific Diagram

Introduction

The heterogeneity gap

❖ 이질성 문제 해결을 위한 접근법 - 분리 학습 (Disentangled Learning)

- 모달리티를 공유 표현(Modality-Invariant)과 고유 표현(Modality-Specific)으로 분리



-Pixel Data Diagram of Abraham Lincoln 166 | Download Scientific Diagram

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis

- 2020년 ACM Multimedia에 게재된 논문 (2025년 9월 기준 1,014회 인용)
- 특징을 '공유'와 '고유' 두 개의 부분 공간으로 분리해 이질성 문제를 해결하고자 함.

MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis

Devamanyu Hazarika
School of Computing
National University of Singapore
hazarika@comp.nus.edu.sg

Roger Zimmermann
School of Computing
National University of Singapore
rogerz@comp.nus.edu.sg

Soujanya Poria
ISTD, Singapore University of Technology and Design
sporia@sutd.edu.sg

ABSTRACT

Multimodal Sentiment Analysis is an active area of research that leverages multimodal signals for affective understanding of user-generated videos. The predominant approach, addressing this task, has been to develop sophisticated fusion techniques. However, the heterogeneous nature of the signals creates distributional modality gaps that pose significant challenges. In this paper, we aim to learn effective modality representations to aid the process of fusion. We propose a novel framework, MISA, which projects each modality to two distinct subspaces. The first subspace is modality-invariant, where the representations across modalities learn their commonalities and reduce the modality gap. The second subspace is modality-specific, which is private to each modality and captures their characteristic features. These representations provide a holistic view of the multimodal data, which is used for fusion that leads to task predictions. Our experiments on popular sentiment

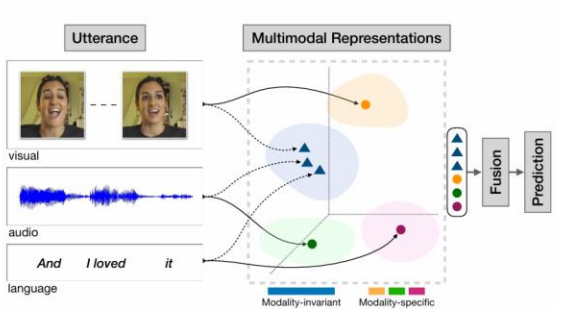


Figure 1: Learning multimodal representations through modality-invariant and -specific subspaces. These features are later utilized for fusion and subsequent prediction of affect in the video.

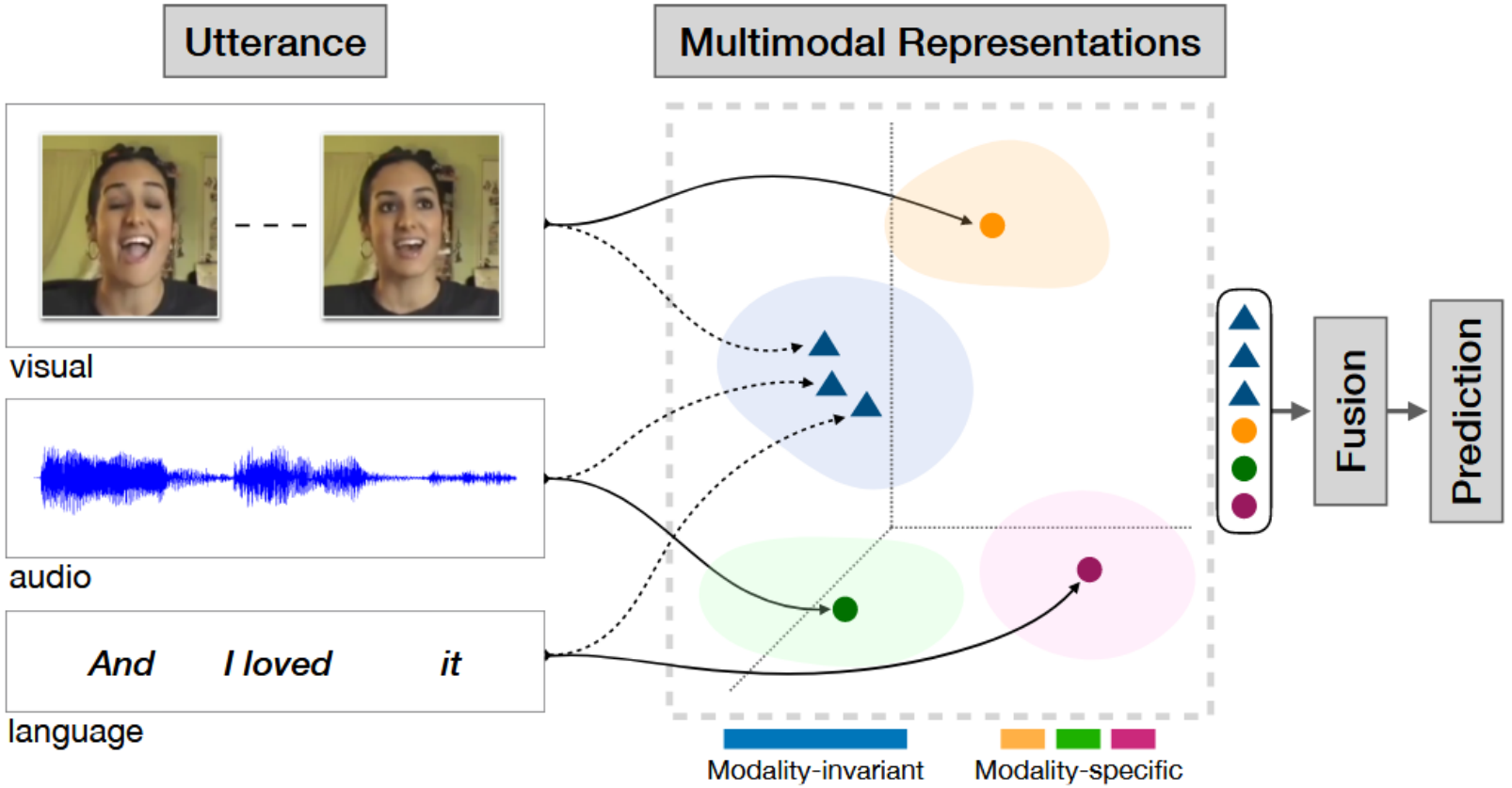
Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ 핵심 아이디어: 표현의 분리

- 각 모달리티의 정보를 두 개의 독립적인 부분 공간(Subspace)로 분리하자

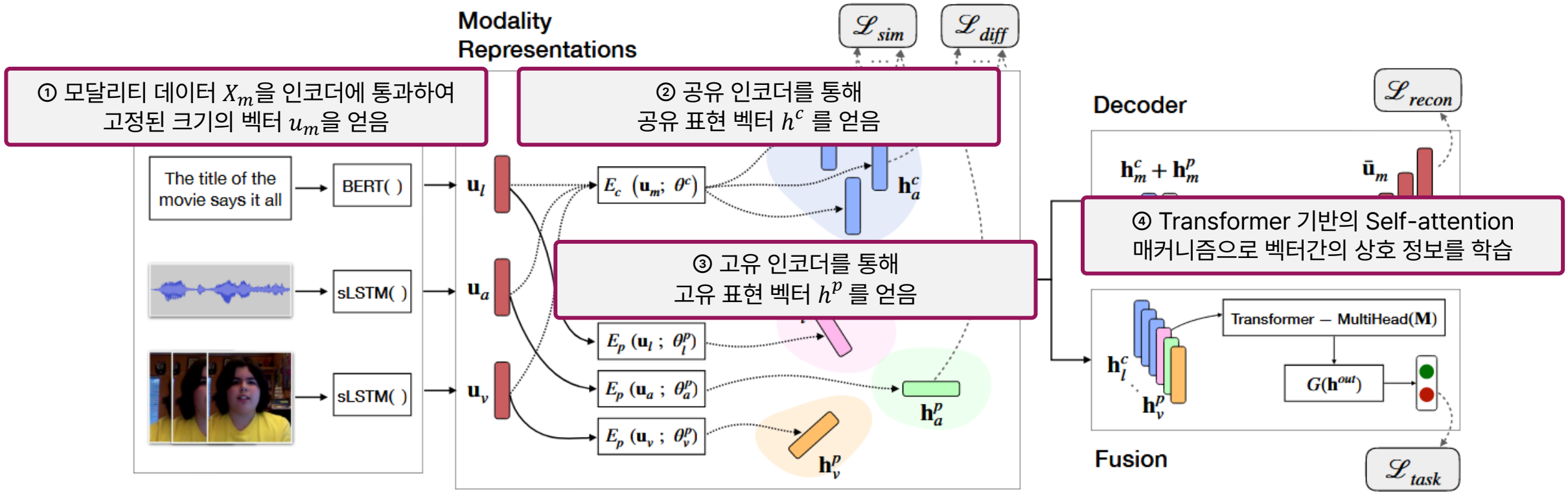


Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ Model Overview



Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ 손실함수

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{sim}} + \beta \mathcal{L}_{\text{diff}} + \gamma \mathcal{L}_{\text{recon}}$$

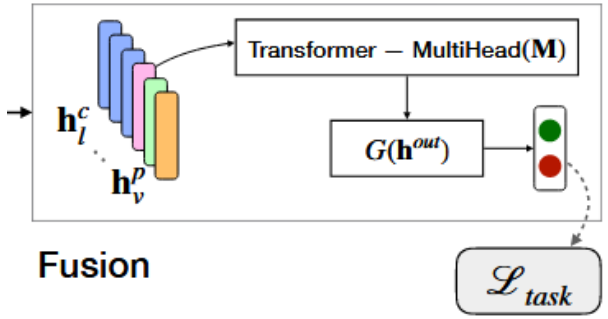
Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ Task Loss

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{sim}} + \beta \mathcal{L}_{\text{diff}} + \gamma \mathcal{L}_{\text{recon}}$$



$$\mathcal{L}_{\text{task}} = -\frac{1}{N_b} \sum_{i=0}^{N_b} \mathbf{y}_i \cdot \log \hat{\mathbf{y}}_i \qquad \text{for classification} \qquad (12)$$

$$= \frac{1}{N_b} \sum_{i=0}^{N_b} \|\mathbf{y}_i - \hat{\mathbf{y}}_i\|_2^2 \qquad \text{for regression} \qquad (13)$$

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

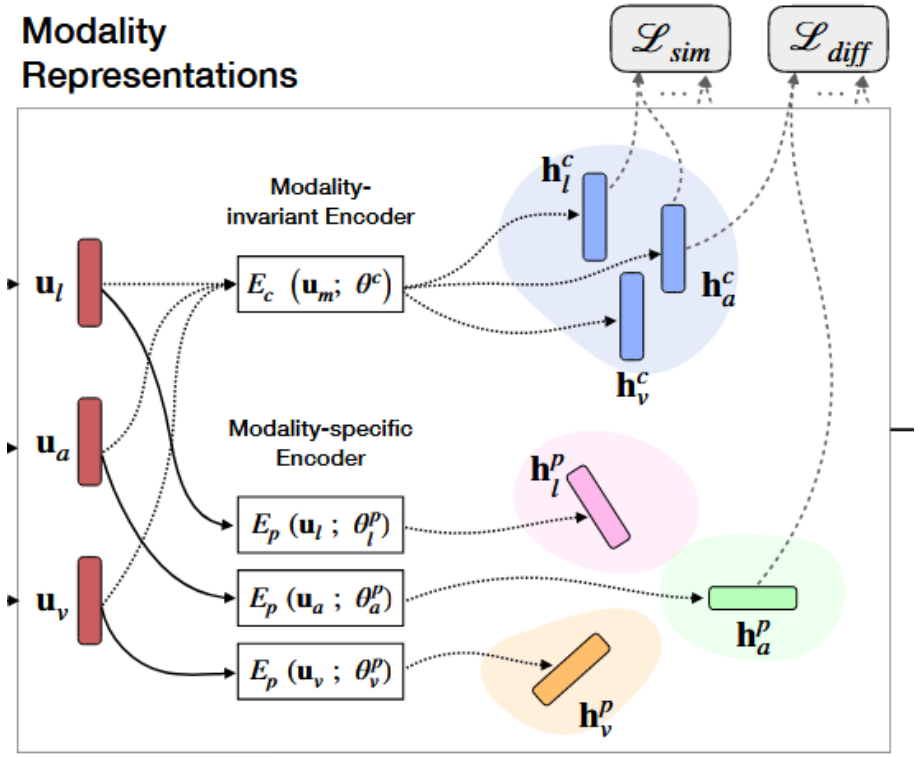
MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ Similarity Loss

- 공유 표현을 공유 공간 상의 한 지점으로 모아주는 역할

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{sim}} + \beta \mathcal{L}_{\text{diff}} + \gamma \mathcal{L}_{\text{recon}}$$



* CMD (Central Moment Discrepancy)
두 데이터 표현의 분포가 얼마나 다른지 측정하는 지표

$$\mathcal{L}_{\text{sim}} = \frac{1}{3} \sum_{\substack{(m_1,m_2) \in \\ \{(l,a),(l,v),\\(a,v)\}}} CMD_K(\mathbf{h}_{m_1}^c, \mathbf{h}_{m_2}^c)$$

$$CMD_K(X,Y) = \frac{1}{|b-a|} \|\mathbf{E}(X) - \mathbf{E}(Y)\|_2 + \sum_{k=2}^{K} \frac{1}{|b-a|^k} \|C_k(X) - C_k(Y)\|_2$$

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

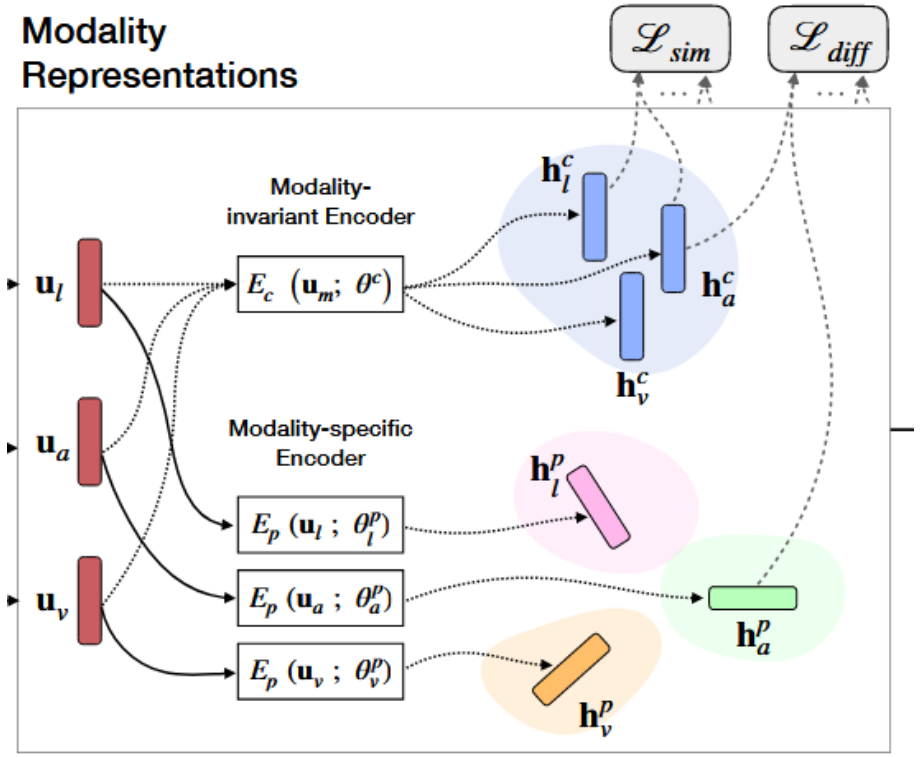
MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ Difference Loss

- 공유 표현과 고유 표현이 서로 다른 정보를 포착하도록 보장

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\,\mathcal{L}_{\text{sim}} + \beta\,\mathcal{L}_{\text{diff}} + \gamma\,\mathcal{L}_{\text{recon}}$$



① 공유 표현과, 고유 표현을 직교하도록

$$\mathcal{L}_{\text{diff}} = \sum_{m \in \{l,v,a\}} \left\| \mathbf{H}_m^{c\top} \mathbf{H}_m^p \right\|_F^2 + \sum_{\substack{(m_1,m_2) \in \\ \{(l,a),(l,v),\\(a,v)\}}} \left\| \mathbf{H}_{m_1}^{p\top} \mathbf{H}_{m_2}^p \right\|_F^2$$

② 모달리티간 고유 표현끼리
서로 직교하도록

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

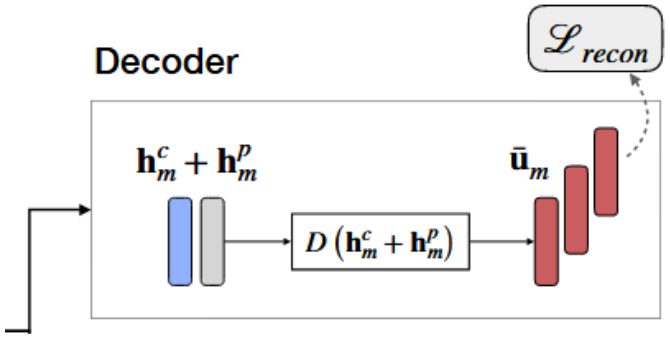
Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ Reconstruction Loss

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathcal{L}_{\text{sim}} + \beta \mathcal{L}_{\text{diff}} + \gamma \mathcal{L}_{\text{recon}}$$

- 학습 시 직교 조건만을 만족시키는 트릭을 사용하지 않도록 하는 안전장치 역할

① 분리했던 공유 표현과 고유 표현을 다시 더함



② Decoder로 원래의 입력 벡터 u_m을 복원하도록 시도

$$\mathcal{L}_{\text{recon}} = \frac{1}{3}\left(\sum_{m\in\{l,v,a\}} \frac{\|\mathbf{u}_m - \hat{\mathbf{u}}_m\|_2^2}{d_h}\right)$$

③ 원래의 입력 벡터 u_m과 복원된 벡터 $\hat{u}_m$의 MSE

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ 실험 환경

- 두 가지 과제에 대한 벤치마크 데이터셋을 사용

❖ Multimodal Sentiment Analysis (MSA)

- MOSI, MOSEI 데이터셋 (유튜브 영상 속 발화)
- -3(강한 부정)부터 +3(강한 긍정)까지의 연속적인 감성 점수가 라벨링 되어있음

❖ Multimodal Humor Detection (MHD)

- UR-FUNNY 데이터셋 (TED 강연의 유머러스한 발화)
- 유머/비유머 이진 라벨

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ 성능 평가 (Multimodal Sentiment Analysis)

- 비교 방법론과 비교했을 때 SOTA 달성
- 복잡한 TFN이나 LMF 같은 모델보다 우수한 성능을 보임

좌측: 부정 vs. 부정이 아닌 것
우측: 부정 vs. 긍정

Models	MOSEI				
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-Score (↑)	Acc-7 (↑)
MFN$^{\otimes}$	-	-	76.0 / -	76.0 / -	-
MV-LSTM$^{\otimes}$	-	-	76.4 / -	76.4 / -	-
Graph-MFN$^{\otimes}$	0.710	0.540	76.9 / -	77.0 / -	45.0
RAVEN	0.614	0.662	79.1 / -	79.5 / -	50.0
MCTN	0.609	0.670	79.8 / -	80.6 / -	49.6
CIA	0.680	0.590	80.4 / -	78.2 / -	50.1
CIM-MTL	-	-	80.5 / -	78.8 / -	-
DFF-ATMF (B)	-	-	- / 77.1	- / 78.3	-
MulT	0.580	0.703	- / 82.5	- / 82.3	51.8
TFN (B)$^{\diamond}$	0.593	0.700	- / 82.5	- / 82.1	50.2
LMF (B)$^{\diamond}$	0.623	0.677	- / 82.0	- / 82.1	48.0
MFM (B)$^{\diamond}$	0.568	0.717	- / 84.4	- / 84.3	51.3
ICCN (B)	0.565	0.713	- / 84.2	- / 84.2	51.6
MISA (B)	**0.555**	**0.756**	**83.6**† / **85.5**†	**83.8 / 85.3**	**52.2**
Δ_{SOTA}	↓ 0.010	↑ 0.043	↑ 3.1 / ↑ 1.3	↑ 5.0 / ↑ 1.1	↑ 0.6

Models	MOSI				
	MAE (↓)	Corr (↑)	Acc-2 (↑)	F-Score (↑)	Acc-7 (↑)
BC-LSTM	1.079	0.581	73.9 / -	73.9 / -	28.7
MV-LSTM	1.019	0.601	73.9 / -	74.0 / -	33.2
TFN	0.970	0.633	73.9 / -	73.4 / -	32.1
MARN	0.968	0.625	77.1 / -	77.0 / -	34.7
MFN	0.965	0.632	77.4 / -	77.3 / -	34.1
LMF	0.912	0.668	76.4 / -	75.7 / -	32.8
CH-Fusion	-	-	80.0 / -	-	-
MFM$^{\otimes}$	0.951	0.662	78.1 / -	78.1 / -	36.2
RAVEN$^{\otimes}$	0.915	0.691	78.0 / -	76.6 / -	33.2
RMFN$^{\otimes}$	0.922	0.681	78.4 / -	78.0 / -	38.3
MCTN$^{\otimes}$	0.909	0.676	79.3 / -	79.1 / -	35.6
CIA	0.914	0.689	79.8 / -	- / 79.5	38.9
HFFN$^{\oslash}$	-	-	- / 80.2	- / 80.3	-
LMFN$^{\oslash}$	-	-	- / 80.9	- / 80.9	-
DFF-ATMF (B)	-	-	- / 80.9	- / 81.2	-
ARGF	-	-	- / 81.4	- / 81.5	-
MulT	0.871	0.698	- / 83.0	- / 82.8	40.0
TFN (B)$^{\diamond}$	0.901	0.698	- / 80.8	- / 80.7	34.9
LMF (B)$^{\diamond}$	0.917	0.695	- / 82.5	- / 82.4	33.2
MFM (B)$^{\diamond}$	0.877	0.706	- / 81.7	- / 81.6	35.4
ICCN (B)	0.860	0.710	- / 83.0	- / 83.0	39.0
MISA (B)	**0.783**	**0.761**	**81.8**† / **83.4**†	**81.7 / 83.6**	**42.3**
Δ_{SOTA}	↓ 0.077	↑ 0.051	↑ 2.0 / ↑ 0.4	↑ 2.6 / ↑ 0.6	↑ 3.3

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ 성능 평가 (Multimodal Humor Detection)

- 모델의 일반화 능력을 검증

Algorithms	context	target	UR_FUNNY Accuracy-2 (↑)
C-MFN	✓		58.45
C-MFN		✓	64.47
TFN		✓	64.71
LMF		✓	65.16
C-MFN	✓	✓	65.23
LMF (Bert)		✓	67.53
TFN (Bert)		✓	68.57
MISA (GloVe)		✓	68.60
MISA (Bert)		✓	**70.61**†
Δ_{SOTA}			↑ 2.07

BERT를 이용한 다른 모델의 성능보다
GloVe를 이용한 MISA의 성능이 우수

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

MISA

Modality-Invariant and Specific Representations for Multimodal Sentiment Analysis (2020, ACM Multimedia)

❖ 공유와 고유라는 이진 분리 프레임워크를 새롭게 제시

❖ But, 모든 모달리티가 Task에 동등하게 기여하며, 그들의 표현 능력이 비슷할 것이라는 가정
→ 현실에서는 모달리티 별 성능이 동등하지 않음

Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis (No. arXiv:2005.03545). arXiv. https://doi.org/10.48550/arXiv.2005.03545

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ Decoupled Multimodal Distilling for Emotion Recognition

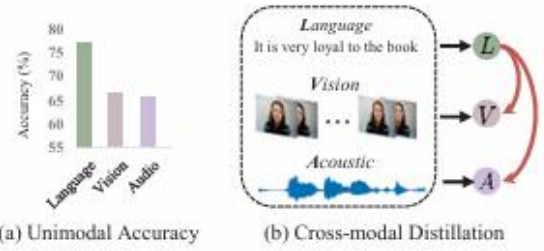
- 2023년 CVPR에 발표된 논문 (2025년 9월 기준 190회 인용)

Decoupled Multimodal Distilling for Emotion Recognition

Yong Li, Yuanzhi Wang, Zhen Cui*
PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China.
{yong.li, yuanzhiwang, zhen.cui}@njust.edu.cn

Abstract

Human multimodal emotion recognition (MER) aims to perceive human emotions via language, visual and acoustic modalities. Despite the impressive performance of previous MER approaches, the inherent multimodal heterogeneities still haunt and the contribution of different modalities varies significantly. In this work, we mitigate this issue



(a) Unimodal Accuracy (b) Cross-modal Distillation

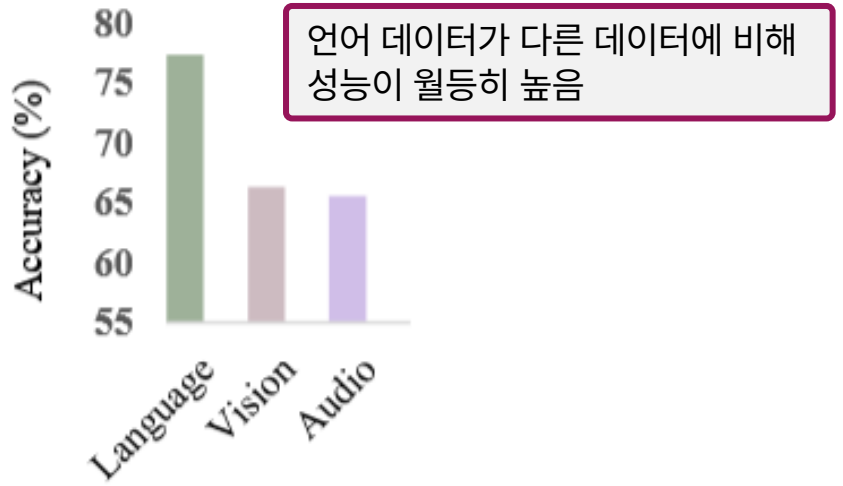
Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

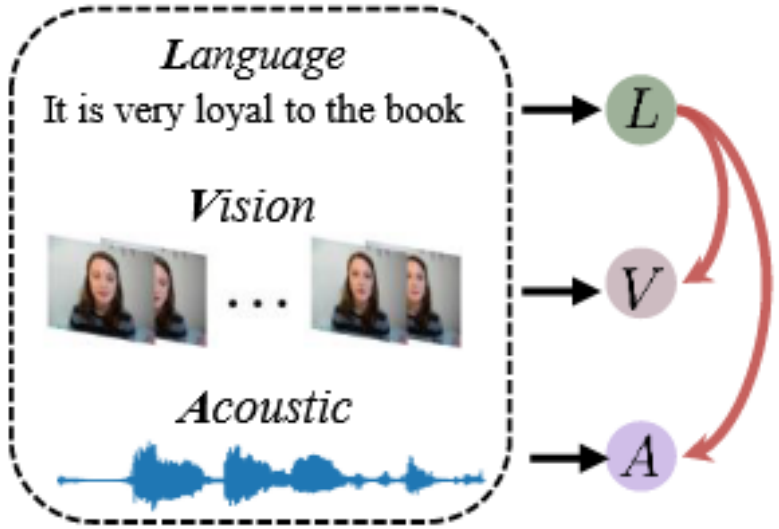
❖ 문제: 데이터 간의 성능 불균형

- 단순히 합치기만 해서는 상대적으로 약한 모달리티의 정보가 제대로 활용되지 못할 것임.
- 지식 증류(Knowledge Distillation) 개념 도입
  - 성능이 좋은 모델의 지식을 성능이 낮은 모델에게 전달



언어 데이터가 다른 데이터에 비해 성능이 월등히 높음

(a) Unimodal Accuracy



언어의 지식을 시각과 음성에 전달

(b) Cross-modal Distillation

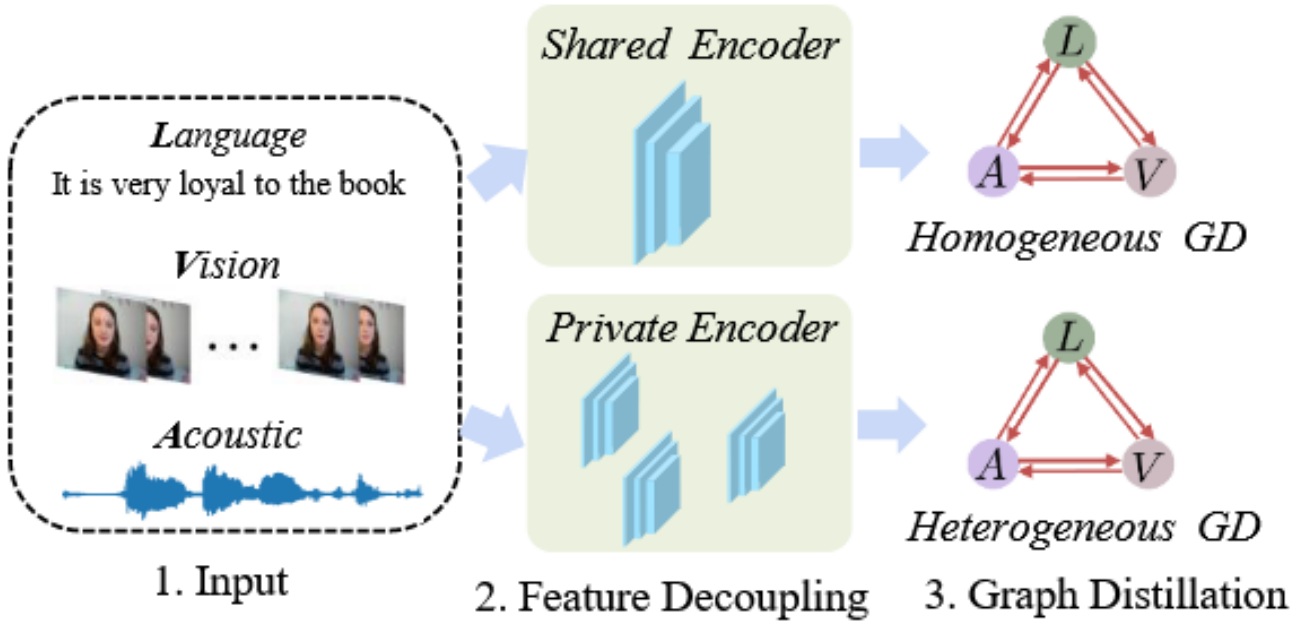
Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ 핵심 아이디어

- Feature Decoupling – 각 데이터의 특징을 공유(Shared) 특징과 고유(Private) 특징으로 분리
- Graph Distillation – 분리된 두 공간에서 모달리티간 지식 전달



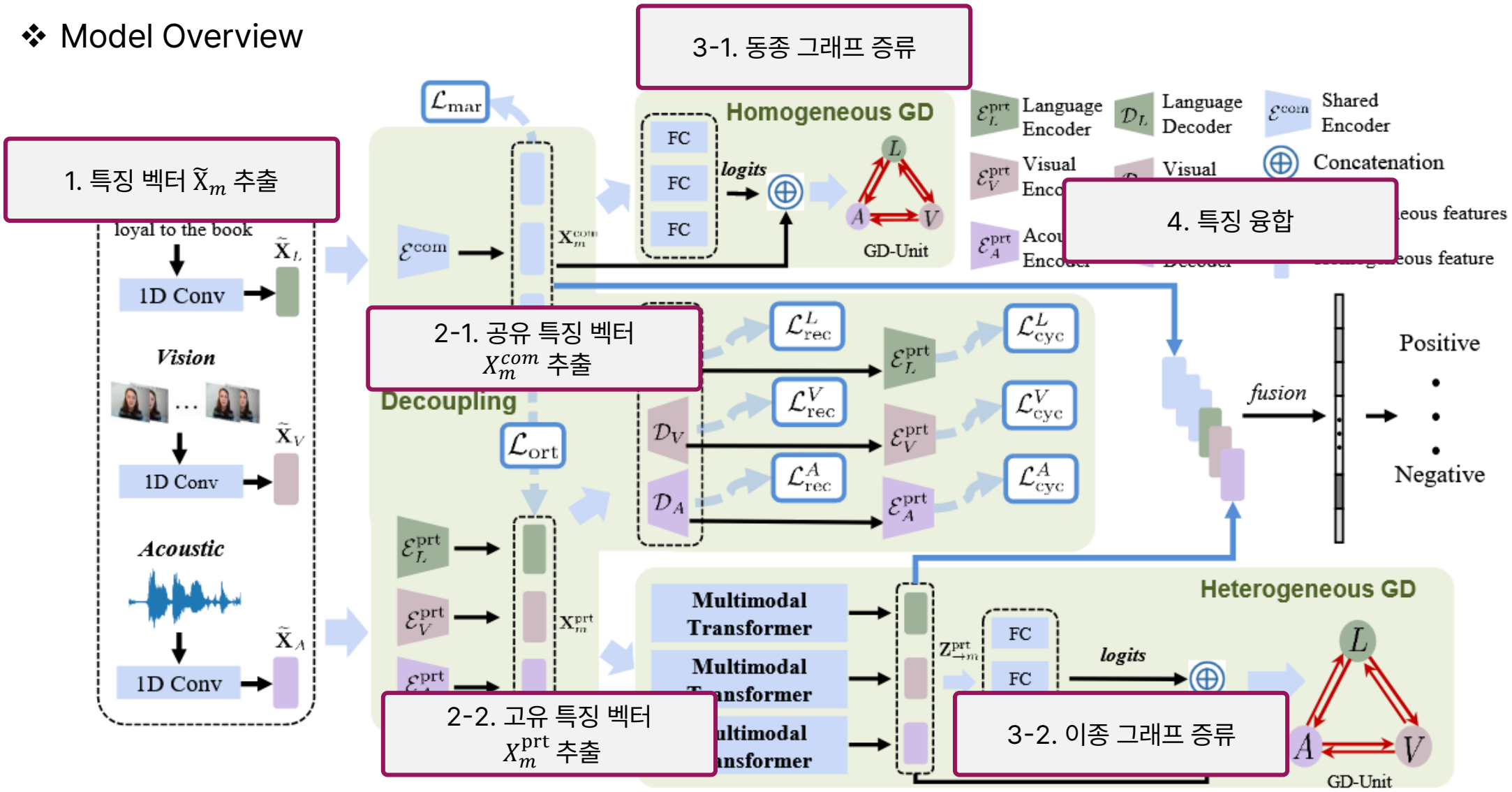
(c) Our proposed Decoupled Multimodal Distillation

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ Model Overview



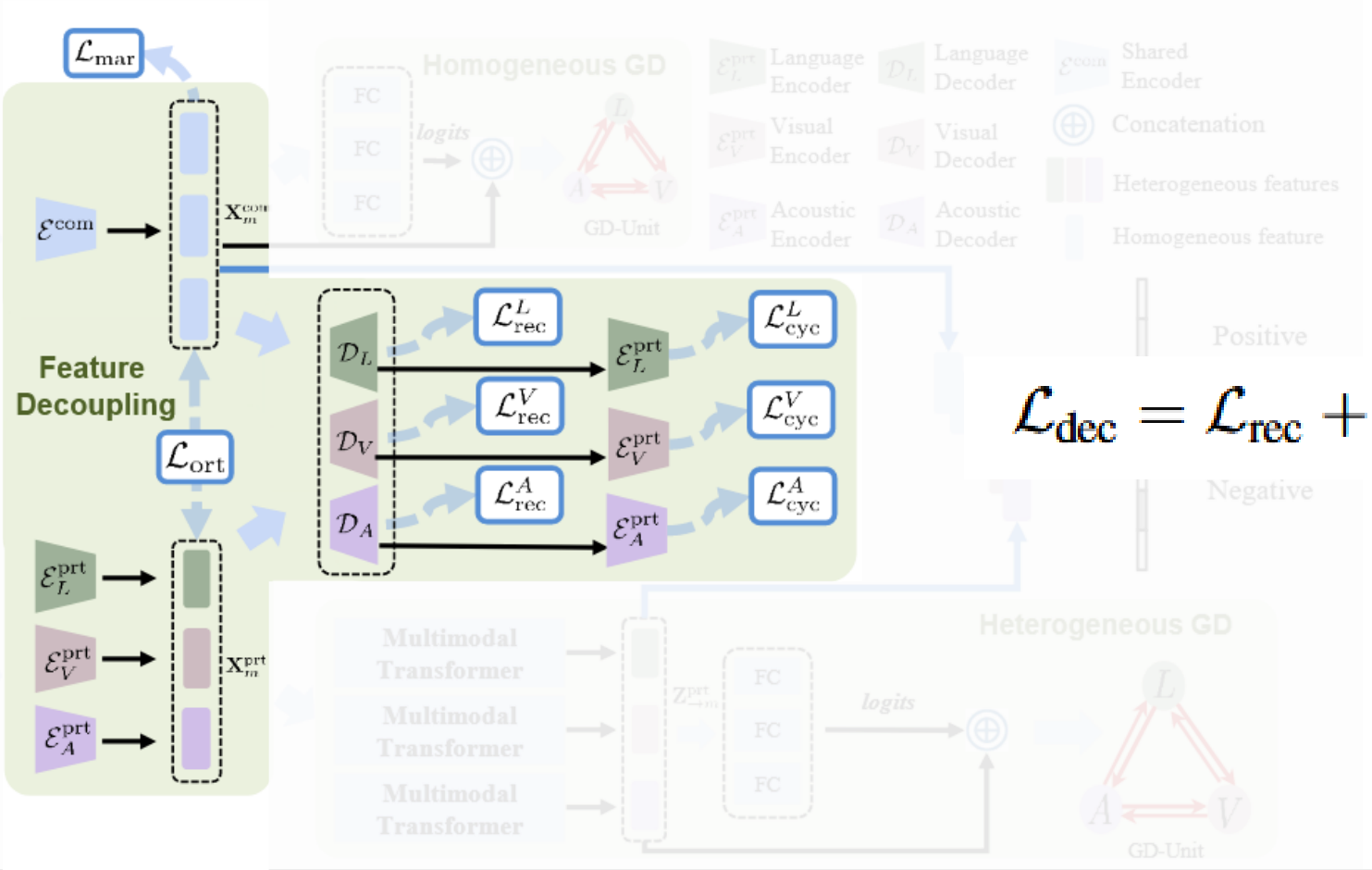
Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ Feature Decoupling Module

- 정교한 분리를 위해 세 가지 손실 함수를 사용



$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}} + \gamma(\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{ort}})$$

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

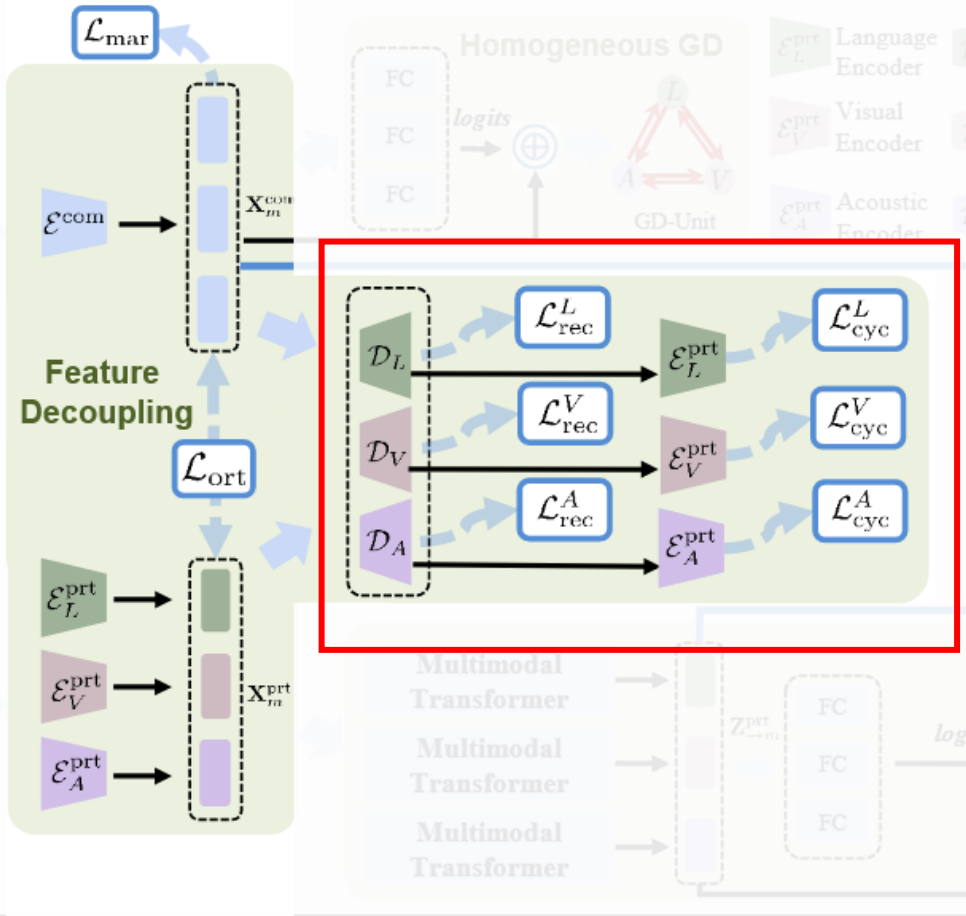
DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ Self-Regression Loss

- 분리된 두 특징이 원본의 정보를 담고 있도록 학습

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}} + \gamma(\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{ort}})$$



$$\mathcal{L}_{\text{rec}} = \|\widetilde{\mathbf{X}}_m - \mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}])\|_F^2$$

$$\mathcal{L}_{\text{cyc}} = \|\mathbf{X}_m^{\text{prt}} - \mathcal{E}_m^{\text{prt}}(\mathcal{D}_m([\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}]))\|_F^2$$

모델이 모든 정보를 개별 특징에만 담아 놓는 편법을 방지하기 위해 추가

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

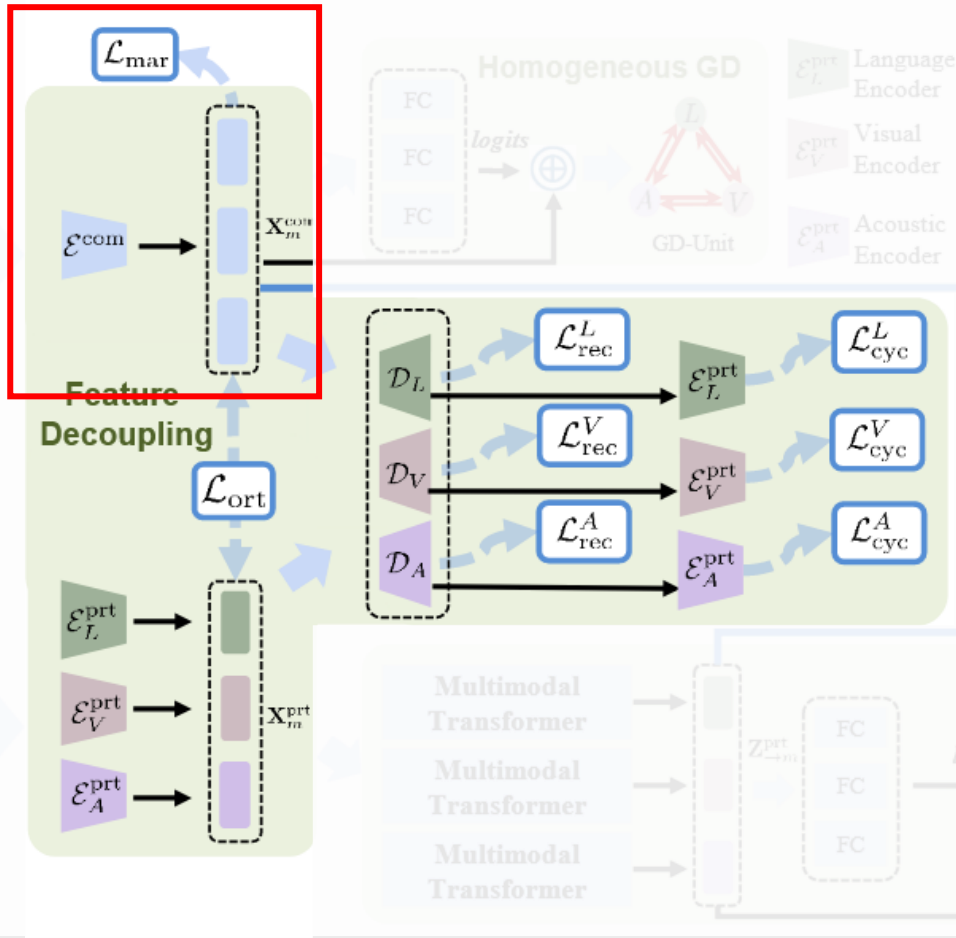
DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ Margin Loss

- 공유 특징 벡터가 같은 의미를 갖도록 학습

$$\mathcal{L}_{dec} = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \gamma(\mathcal{L}_{mar} + \mathcal{L}_{ort})$$



가설: 공통 특징 공간에서, 같은 감정을 나타내는 샘플이라면 어떤 모달리티에서 왔든 관계없이 유사해야 함.

샘플 i 와 j 는 다른 모달리티에서 온 같은 감정을 갖는 긍정쌍
샘플 i 와 k 는 같은 모달리티에서 온 다른 감정을 갖는 부정쌍

$$S = \left\{(i,j,k) \middle| \begin{matrix} m[i]\neq m[j],\ m[i]=m[k] \\ c[i]=c[j],\ c[i]\neq c[k] \end{matrix}\right\}$$

긍정쌍 (i,j)는 유사하도록,
부정쌍 (i,k)는 멀어지도록

$$\mathcal{L}_{mar} = \frac{1}{|S|}\sum_{(i,j,k)\in S} \max(0, \alpha - \cos(X_{m[i]}^{com}, X_{m[j]}^{com}) + \cos(X_{m[i]}^{com}, X_{m[k]}^{com}))$$

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

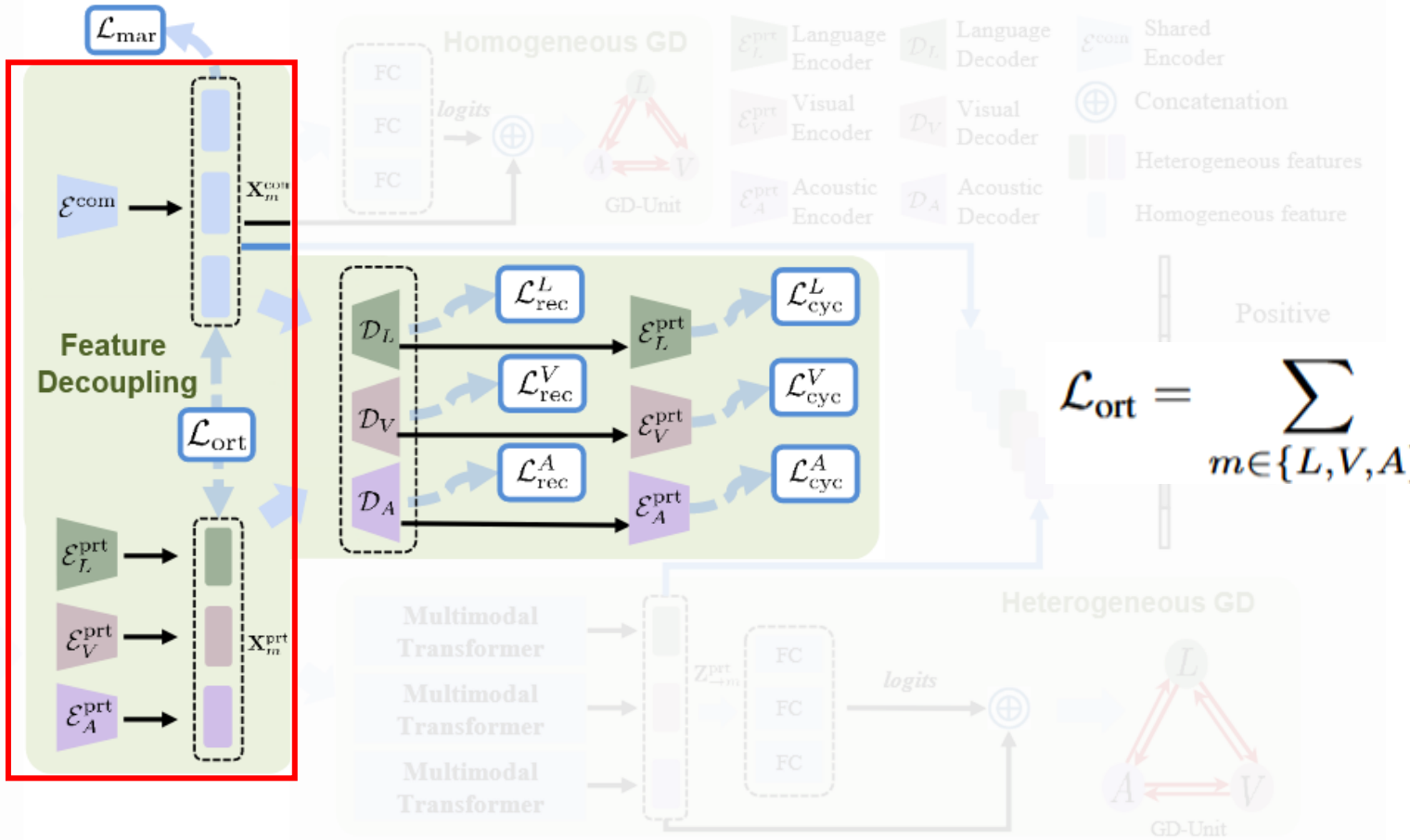
DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ Orthogonality Loss

- 공유 특징과 고유 특징이 직교하도록 학습

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{cyc}} + \gamma(\mathcal{L}_{\text{mar}} + \mathcal{L}_{\text{ort}})$$



$$\mathcal{L}_{\text{ort}} = \sum_{m \in \{L,V,A\}} \cos(\mathbf{X}_m^{\text{com}}, \mathbf{X}_m^{\text{prt}}).$$

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

분리 손실 $\mathcal{L}_{\text{dec}} = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \gamma(\mathcal{L}_{mar} + \mathcal{L}_{ort})$

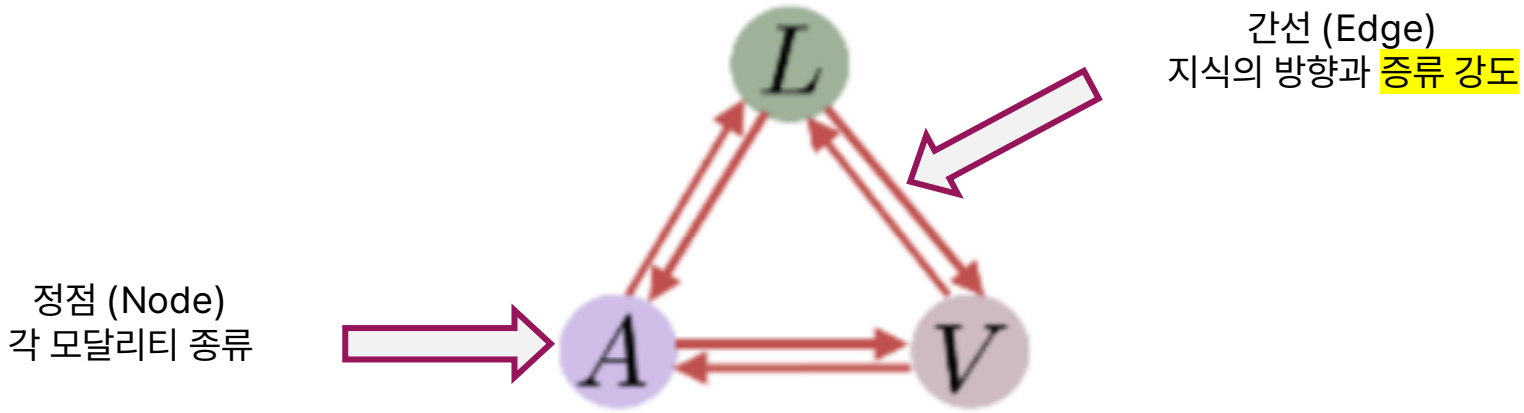
Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ GD-Unit

- 약한 성능의 모달리티 정보를 활용하지 못하는 문제를 해결하기 위해 지식 증류(Knowledge Distillation)를 도입
- 그러나 멀티모달 데이터는 어떤 모달리티가 강한 모달리티인지 알기 어려움.
- 모든 모달리티가 서로에게 지식을 전달하고 지식의 강도를 스스로 학습하도록 하는 방식의 GD-Unit(Graph Distillation Unit)을 제안



GD-Unit

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

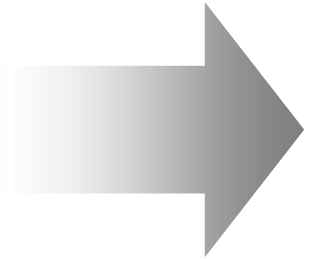
❖ GD-Unit 의 목적

- 증류 오차(E)와 증류 강도(W)를 사용해 모달리티 간의 성능 격차를 줄이는 것

❖ 증류 오차 (E)

- 한 모달리티의 예측값과 다른 모달리티의 예측 값의 차이
- 모달리티 사이의 지식 격차나 예측 불일치를 의미

$$\epsilon_{i\to j} = f(X_i, \theta_i) - f(X_j, \theta_j)$$



$$E = \begin{pmatrix} 0 & \epsilon_{a\to v} & \epsilon_{a\to l} \\ \epsilon_{v\to a} & 0 & \epsilon_{v\to l} \\ \epsilon_{l\to a} & \epsilon_{l\to v} & 0 \end{pmatrix}$$

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ GD-Unit 의 목적

- 증류 오차(E)와 증류 강도(W)를 사용해 모달리티 간의 성능 격차를 줄이는 것

❖ 증류 강도 (W)

- 역할: 다른 모달리티로 지식을 전달할 때의 가중치
- 모달리티의 특징 벡터와 예측 결과 값을 결합해 자동으로 학습함

$$w_{(i\to j)} = g\left(\left[\left[f\left(X_i, \theta_i^1\right), X_i\right], \left[f\left(X_j, \theta_j^1\right), X_j\right]\right], \theta^2\right)$$

$$W = \begin{pmatrix} 0 & w_{a\to v} & w_{a\to l} \\ w_{v\to a} & 0 & w_{v\to l} \\ w_{l\to a} & w_{l\to v} & 0 \end{pmatrix}$$

① 얼마나 이 예측에 자신이 있는가?

② 예측의 근거가 되는 정보가 타당한가?

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ GD-Unit 의 목적

- 증류 오차(E)와 증류 강도(W)를 사용해 모달리티 간의 성능 격차를 줄이는 것

증류 오차

$$E = \begin{pmatrix} 0 & \epsilon_{a\to v} & \epsilon_{a\to l} \\ \epsilon_{v\to a} & 0 & \epsilon_{v\to l} \\ \epsilon_{l\to a} & \epsilon_{l\to v} & 0 \end{pmatrix}$$

증류 강도

$$W = \begin{pmatrix} 0 & w_{a\to v} & w_{a\to l} \\ w_{v\to a} & 0 & w_{v\to l} \\ w_{l\to a} & w_{l\to v} & 0 \end{pmatrix}$$

$$\mathcal{L}_{\mathrm{dtl}} = \|\mathrm{E} \odot W\|_1$$

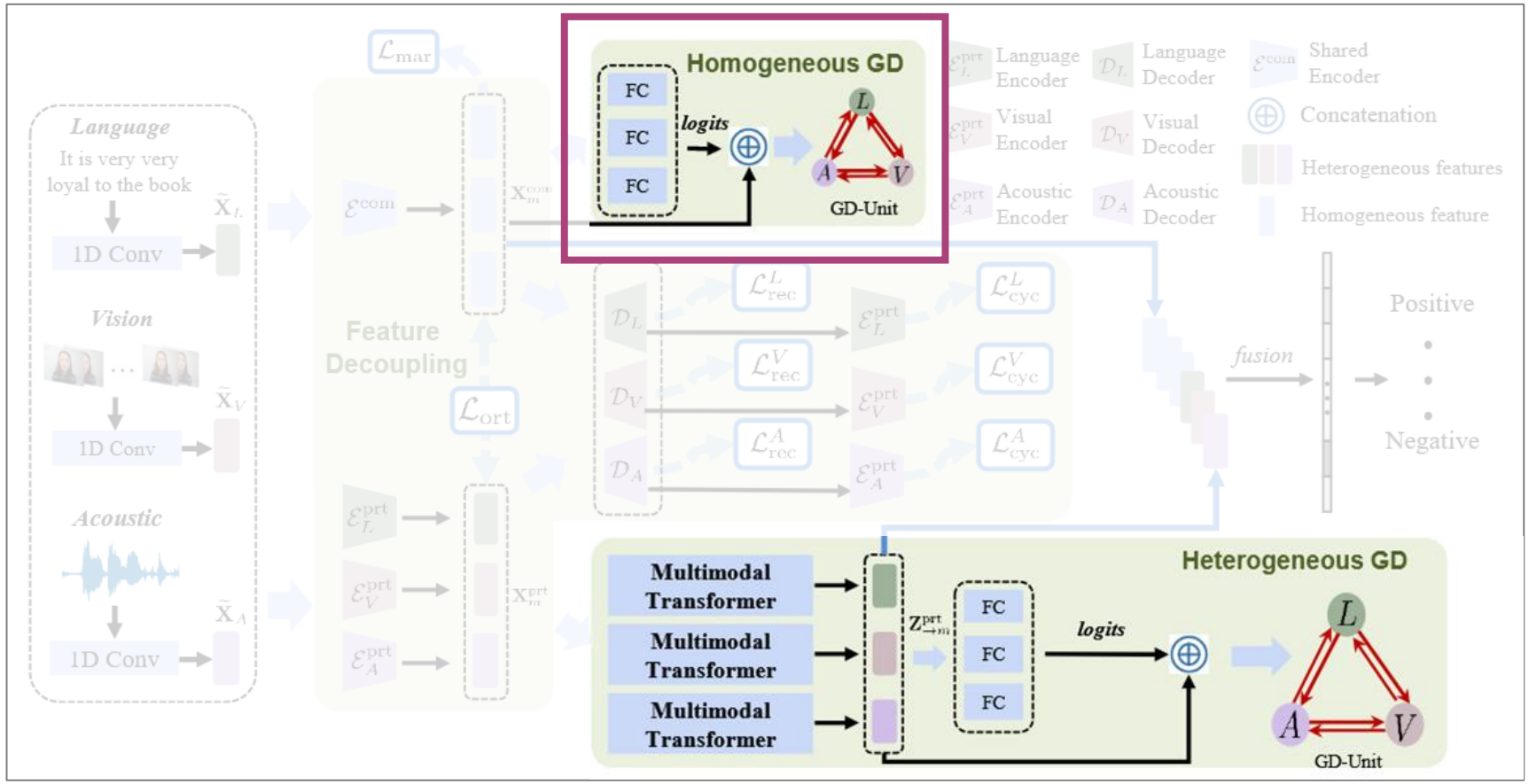
Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ GD-Unit 적용

- GD-Unit을 공통 특징 공간과 고유 특징 공간 두 군데에 적용함



Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

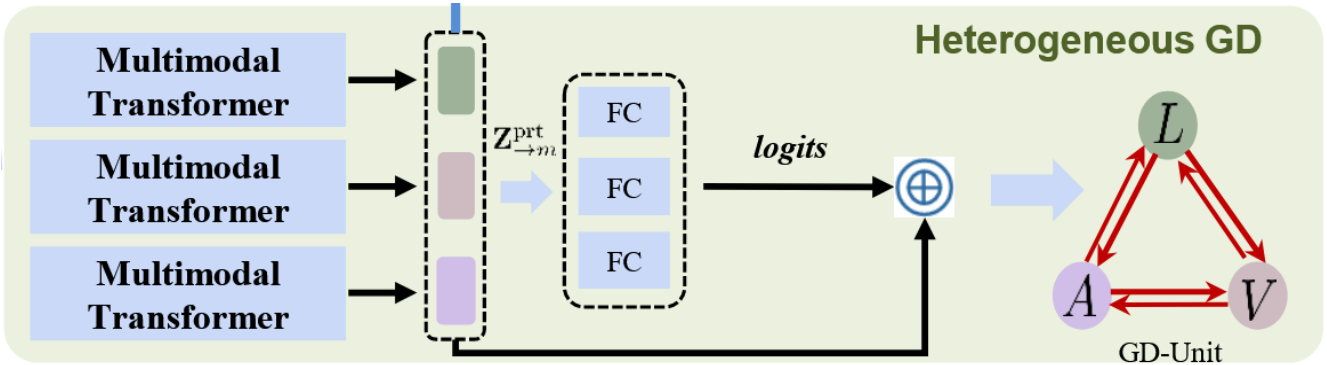
Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ GD-Unit 적용

- GD-Unit을 공통 특징 공간과 고유 특징 공간 두 군데에 적용함

❖ Hetrogeneous GD

- 모달리티간의 분포 차이가 크고 이질적인 특징들
- Multimodal Transformer 및 Cross Attention을 통해 분포 격차를 줄임



Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

증류 손실 $\mathcal{L}_{\text{dtl}} = \mathcal{L}_{dtl}^{homo} + \mathcal{L}_{dtl}^{hetero}$

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ 손실함수

$$\mathcal{L}_{\text{dtl}} = \mathcal{L}_{dtl}^{homo} + \mathcal{L}_{dtl}^{hetero}$$

$$\mathcal{L}_{\text{dec}} = \mathcal{L}_{rec} + \mathcal{L}_{cyc} + \gamma(\mathcal{L}_{mar} + \mathcal{L}_{ort})$$

$$\mathcal{L}_{\text{task}} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{dec}} + \lambda_2 \mathcal{L}_{\text{dtl}}$$

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

❖ 성능 평가

- 두 가지 과제에 대한 벤치마크 데이터셋을 사용, SOTA 달성

Table 1. Comparison on CMU-MOSI dataset. **Bold** is the best.

Methods	Setting	ACC_7 (%)	ACC_2 (%)	F1 (%)
EF-LSTM	Aligned	33.7	75.3	75.2
LF-LSTM		35.3	76.8	76.7
TFN [33]		32.1	73.9	73.4
LMF [14]		32.8	76.4	75.7
MFM [29]		36.2	78.1	78.1
RAVEN [30]		33.2	78.0	76.6
MCTN [26]		35.6	79.3	79.1
MulT [28]		40.0	83.0	82.8
PMR [17]		40.6	83.6	83.4
DMD (**Ours**)		**41.4**	**84.5**	**84.4**
MISA [7]*	Aligned	42.3	83.4	83.6
FDMER [32]*		44.1	84.6	84.7
DMD (**Ours**)*		**45.6**	**86.0**	**86.0**
EF-LSTM	Unaligned	31.0	73.6	74.5
LF-LSTM		33.7	77.6	77.8
RAVEN [30]		31.7	72.7	73.1
MCTN [26]		32.7	75.9	76.4
MulT [28]		39.1	81.1	81.0
PMR [17]		40.6	82.4	82.1
MICA [13]		40.8	82.6	82.7
DMD (**Ours**)		**41.9**	**83.5**	**83.5**

* means the input language features are BERT-based.

Table 2. Comparison on CMU-MOSEI dataset. **Bold** is the best.

Methods	Setting	ACC_7 (%)	ACC_2 (%)	F1 (%)
EF-LSTM	Aligned	47.4	78.2	77.9
LF-LSTM		48.8	80.6	80.6
Graph-MFN [36]		45.0	76.9	77.0
RAVEN [30]		50.0	79.1	79.5
MCTN [26]		49.6	79.8	80.6
MulT [28]		51.8	82.5	82.3
PMR [17]		52.5	83.3	82.6
DMD (**Ours**)		**53.7**	**85.0**	**84.9**
MISA [7]*	Aligned	52.2	85.5	85.3
FDMER [32]*		54.1	86.1	85.8
DMD (**Ours**)*		**54.5**	**86.6**	**86.6**
EF-LSTM	Unaligned	46.3	76.1	75.9
LF-LSTM		48.8	77.5	78.2
RAVEN [30]		45.5	75.4	75.7
MCTN [26]		48.2	79.3	79.7
MulT [28]		50.7	81.6	81.6
PMR [17]		51.8	83.1	82.8
MICA [13]		52.4	83.7	83.3
DMD (**Ours**)		**54.6**	**84.8**	**84.7**

* means the input language features are BERT-based.

Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

DMD

Decoupled Multimodal Distilling for Emotion Recognition (2023, CVPR)

MISA, DMD 모두 공유와 고유, 이중 분리 프레임워크를 사용하고 있는데..

과연 이중 분리 프레임워크가 최선일까?

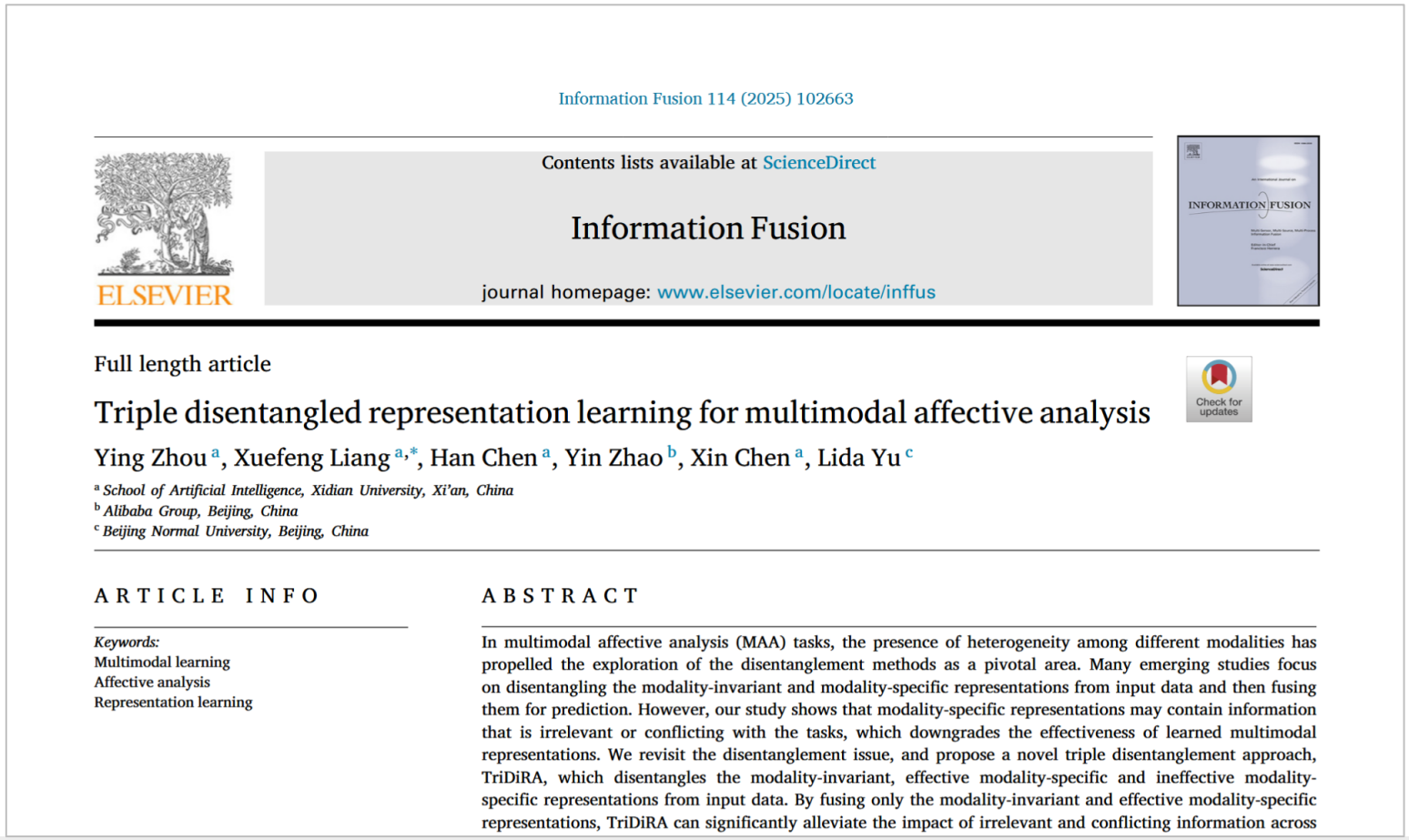
Li, Y., Wang, Y., & Cui, Z. (2023). Decoupled Multimodal Distilling for Emotion Recognition (No. arXiv:2303.13802). arXiv. https://doi.org/10.48550/arXiv.2303.13802

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Triple Disentangled Representation Learning for Multimodal Affective Analysis

- 2025년 Information Fusion에 발표된 논문 (2025년 9월 기준 8회 인용)



Information Fusion 114 (2025) 102663

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Triple disentangled representation learning for multimodal affective analysis

Ying Zhou [a], Xuefeng Liang [a,*], Han Chen [a], Yin Zhao [b], Xin Chen [a], Lida Yu [c]

[a] School of Artificial Intelligence, Xidian University, Xi'an, China
[b] Alibaba Group, Beijing, China
[c] Beijing Normal University, Beijing, China

ARTICLE INFO

Keywords:
Multimodal learning
Affective analysis
Representation learning

ABSTRACT

In multimodal affective analysis (MAA) tasks, the presence of heterogeneity among different modalities has propelled the exploration of the disentanglement methods as a pivotal area. Many emerging studies focus on disentangling the modality-invariant and modality-specific representations from input data and then fusing them for prediction. However, our study shows that modality-specific representations may contain information that is irrelevant or conflicting with the tasks, which downgrades the effectiveness of learned multimodal representations. We revisit the disentanglement issue, and propose a novel triple disentanglement approach, TriDiRA, which disentangles the modality-invariant, effective modality-specific and ineffective modality-specific representations from input data. By fusing only the modality-invariant and effective modality-specific representations, TriDiRA can significantly alleviate the impact of irrelevant and conflicting information across

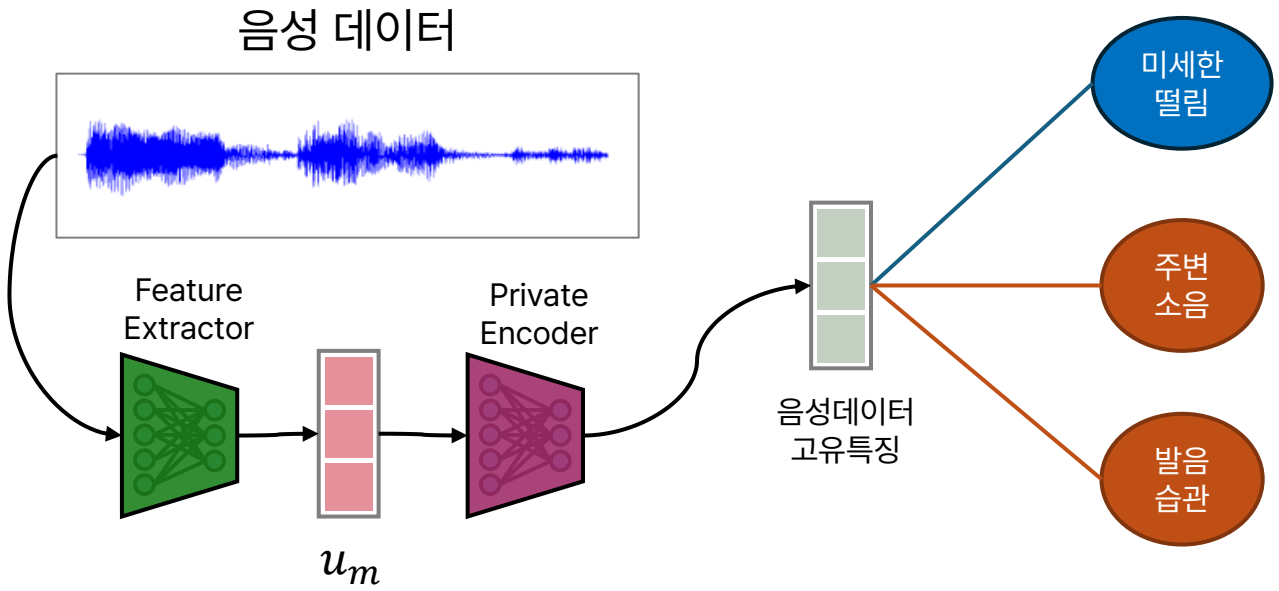
Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 고유 특징이 모두 유용한 정보인가?

- 데이터의 고유 특징 안에는 감정을 드러내는 유용한 정보도 있지만, 주변 소음이나 발음 습관처럼 관련 없는 정보도 존재
- 기존 모델들은 이 둘을 구분 없이 사용하고 있어 모델의 성능을 저해할 수 있음을 강조



고유 특징 벡터에는 유용한 정보,
무관한 정보가 혼재 되어있음

데이터를 삼중 분리하자

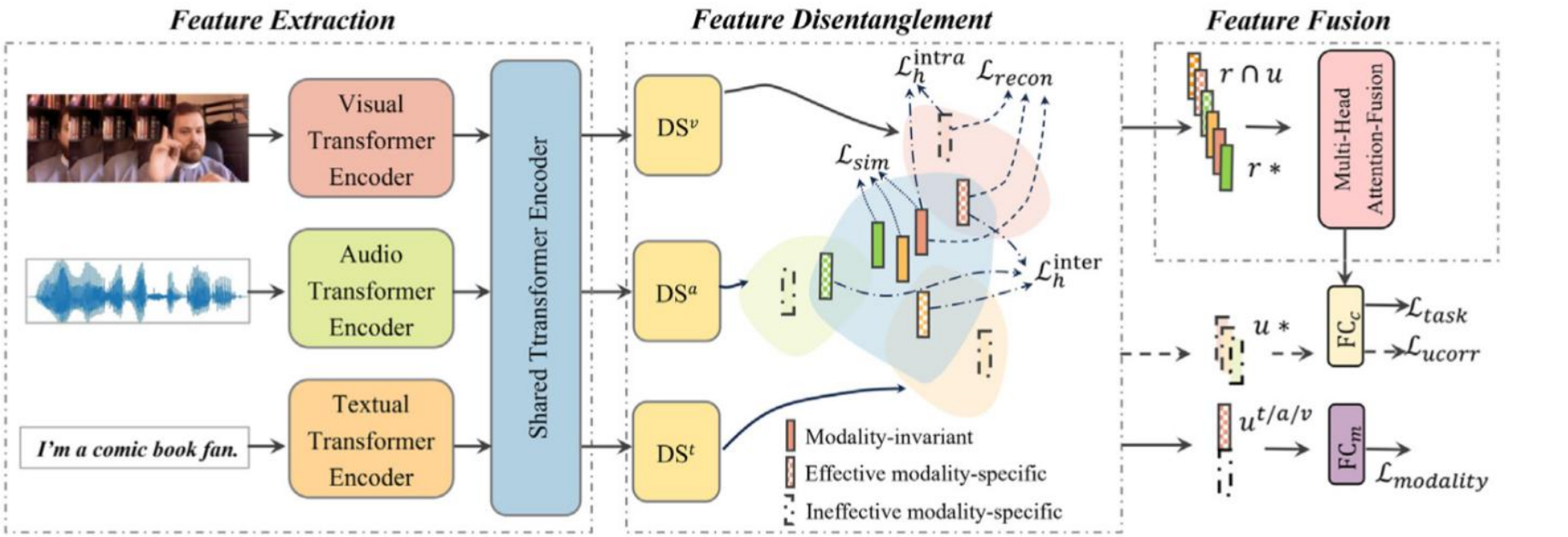
① 공유 특징 (r^*)
② 효과적인 고유 특징 $(r \cap u)$
③ 비효과적인 고유 특징 (u^*)

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Model Overview



① 모달리티의 고유한 특징 보존
② 특징 벡터간 암묵적 정렬 기능

특징 분리
① 공유 특징 (r^*)
② 효과적 고유 특징 ($r \cap u$)
③ 비효과적 고유 특징 (u^*)

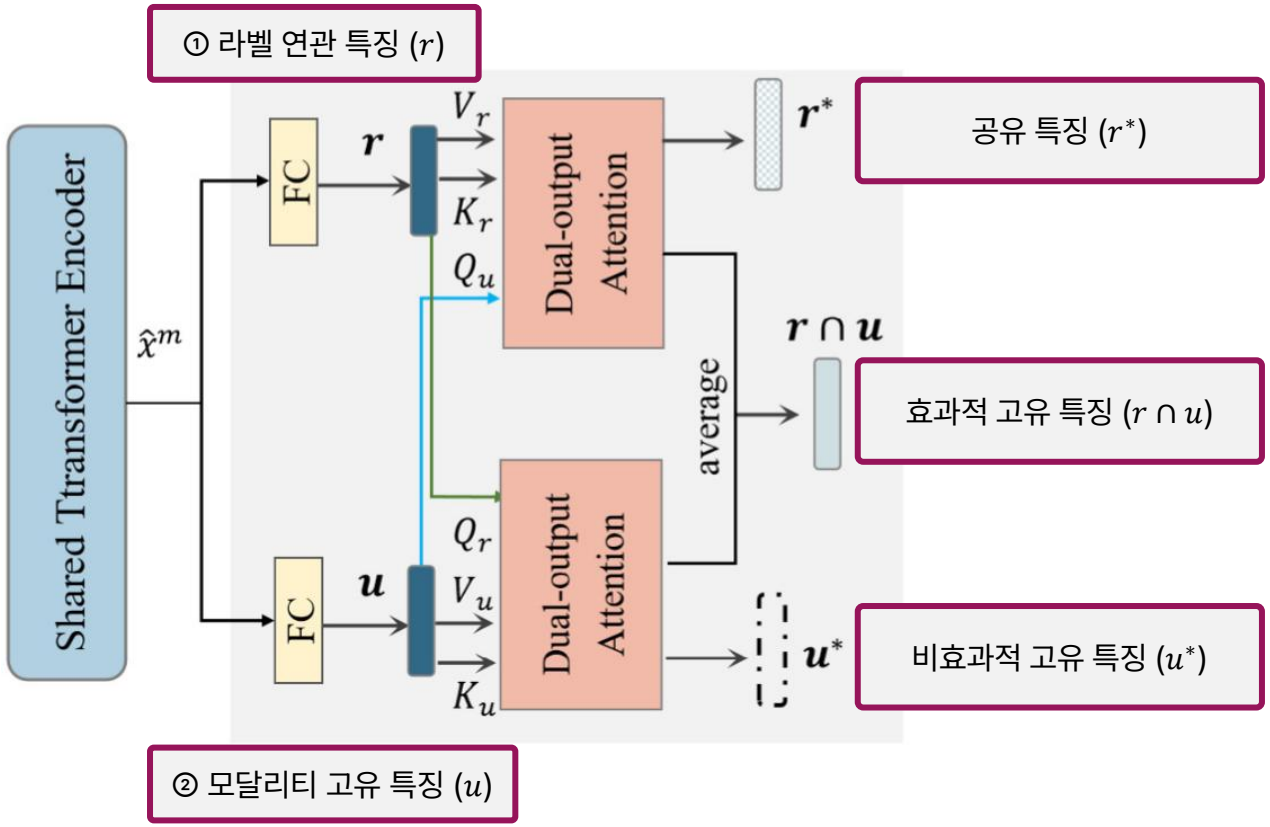
두 가지 특징을 사용해 과제 수행
① 공유 특징 (r^*)
② 효과적 고유 특징 ($r \cap u$)

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

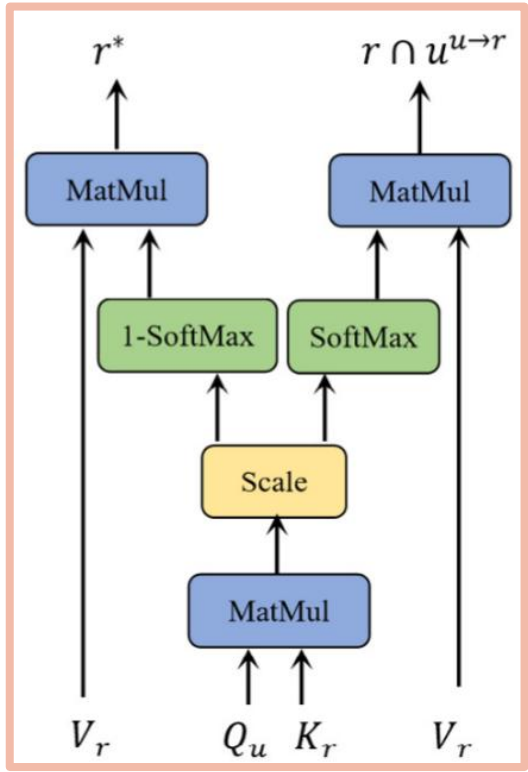
TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Feature Disentanglement



Dual-output Attention

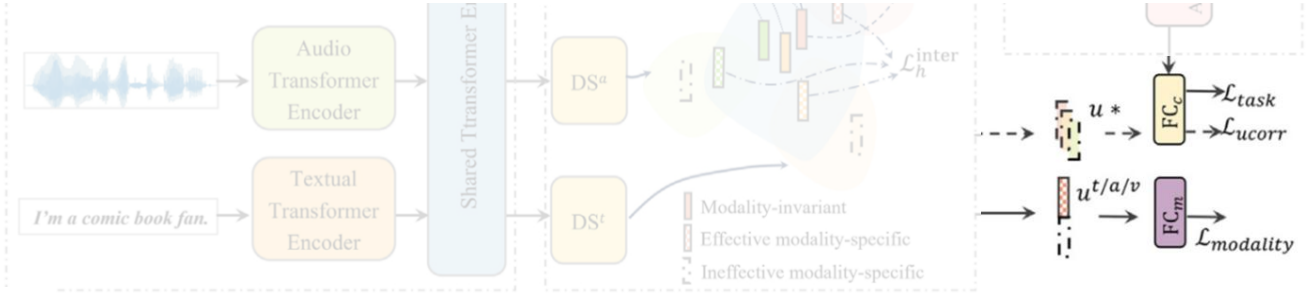


Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 과제 손실



$$\mathcal{L}_{task} = \begin{cases} \frac{1}{B}\sum_{i=1}^{B} \| y_i - \hat{y}_i \|_2^2 & (regression) \\ -\frac{1}{B}\sum_{i=1}^{B} y_i \cdot log\hat{y}_i & (classification). \end{cases}$$

불필요한 고유 특징 $\boldsymbol{u}*$는 예측 과정에서 제외
→ 라벨과 연관된 정보만 담도록 학습됨

$$\mathcal{L}^{m}_{modality} = -\frac{1}{B}\sum_{i=1}^{B} \mathbb{I}(m) \cdot log D(u_i^m),$$

고유 특징들은 모두 모달리티의 고유한 정보를 담고 있어야 함.
→ 고유 특징이 어떤 모달리티에서 왔는지 예측하는 보조 과제 수행

$$\mathcal{L}_{ucorr} = \begin{cases} \frac{cov(\tilde{Y},Y)}{\sigma_{\tilde{Y}}\sigma_Y} & (regression), \\ \frac{1}{B}\sum_{i=1}^{B} y_i \cdot log\tilde{y}_i & (classification). \end{cases}$$

비효과적인 고유 특징(u^*)은 예측 과제와는 관련이 없어야 함.
→ u^*만 사용해서 예측 과제를 수행하고, 실제 정답과 무관하도록 보조 과제 수행

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

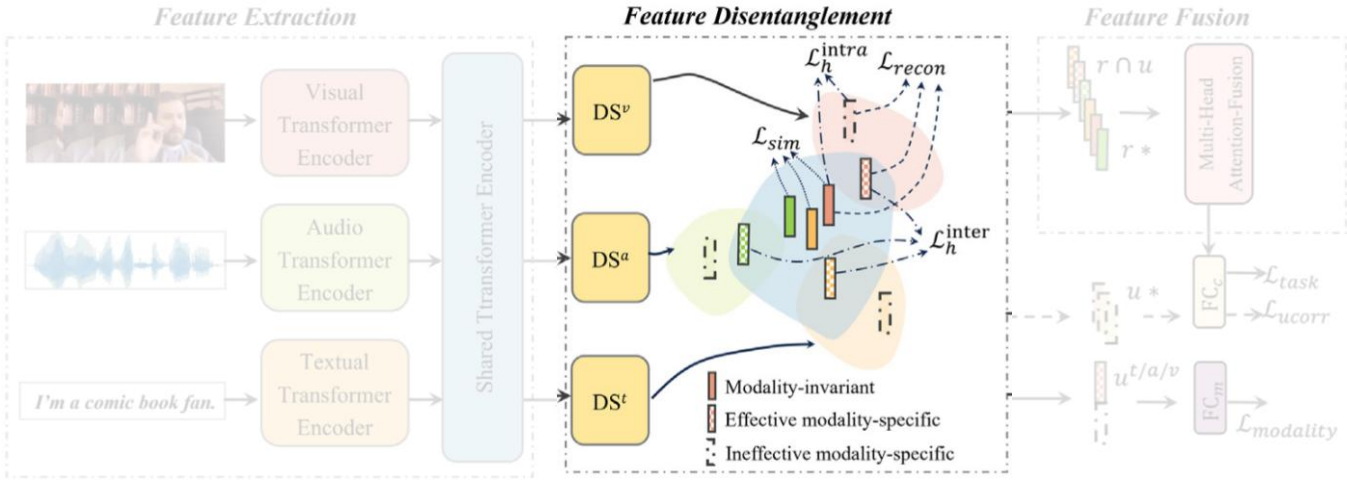
TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 유사성 손실

$$\mathcal{L}_{sim} = \frac{1}{3} \sum_{(m_1,m_2)} CMD_K(r^{*,m_1}, r^{*,m_2}),$$

where $(m_1, m_2) \in (t, a), (t, v), (a, v).$



❖ 독립성 손실

$$\mathcal{L}_h^{inter} = \frac{1}{3} \sum_{(m_1,m_2)} \mathcal{L}_h^{(r\cap u)^{m_1},(r\cap u)^{m_2}}$$

모달리티 간 독립성
"언어의 효과적인 고유 특징"과 "시각의 효과적인 고유 특징"이 독립적인 정보를 담아야 함
→ 각 모달리티의 고유 특징이 섞이지 않도록 분리

$$\mathcal{L}_h^{intra} = \mathcal{L}_h^{r^{*,m},u^{*,m}}$$

모달리티 내 독립성
공유 특징과 비효율적인 고유 특징이 배타적인 특성을 갖도록 학습
→ 공유 특징과 불필요한 노이즈를 분리

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

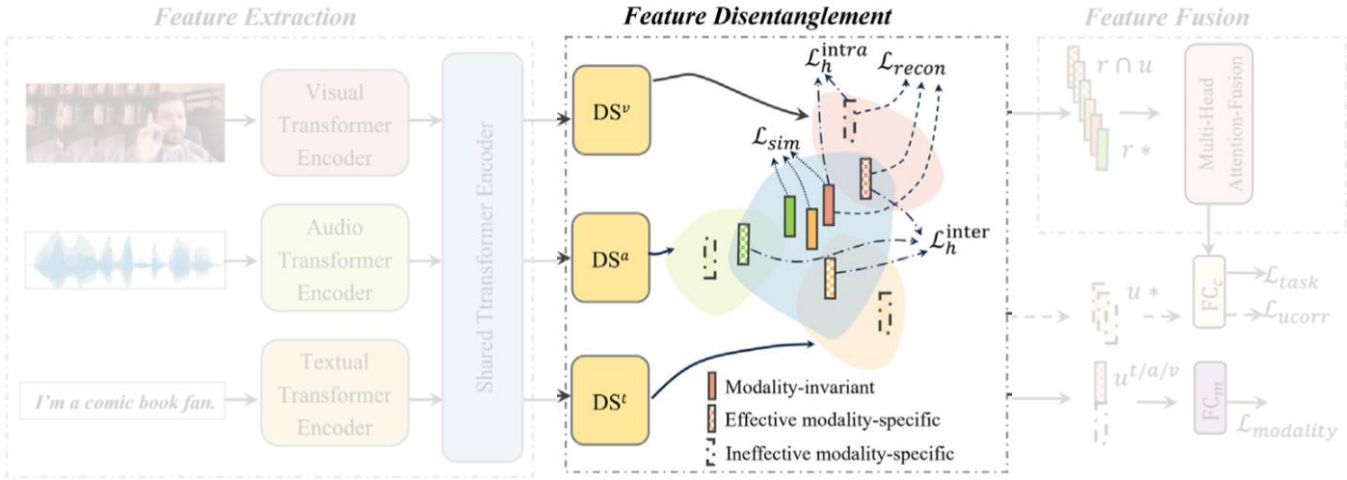
TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 재구성 손실

$$\mathcal{L}_{recon}^{m} = \frac{1}{B}\sum_{i=1}^{B} \|\hat{x}^{m'} - \hat{x}^{m}\|_2^2,$$

$$\hat{x}^{m'} = R[r^{*\mathrm{m}}, (r \cap u)^{m}, u^{*\mathrm{m}}]$$



❖ Total Loss

$$\mathcal{L}_{total} = \mathcal{L}_{\mathrm{task}} + \mathcal{L}_{\mathrm{modality}} + \mathcal{L}_{\mathrm{ucorr}} + \mathcal{L}_{\mathrm{sim}} + \mathcal{L}_{inter} + \mathcal{L}_{intra} + \mathcal{L}_{recon}$$

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 감정 분석 (Multimodal Sentiment Analysis)

- 데이터셋: MOSI, MOSEI
- 평가 지표(회귀): MAE, Corr
- 평가 지표(분류): Acc-7, Acc-2, F1-score

❖ 유머 감지 (Multimodal Humor Detection)

- 데이터셋: UR-FUNNY
- 평가 지표: Acc-2, F1-score

❖ 감정 분류 (Multimodal Emotion Classification)

- 데이터셋: MELD 데이터셋 사용
- 평가 지표: Acc-6

Table 1
The statistics of the datasets.

Dataset	Training set	Valid set	Test set	All
CMU-MOSI	1284	229	686	2199
CMU-MOSEI	16 326	1871	4659	22 856
UR-FUNNY	10 598	2626	3290	16 514
MELD	5280	640	1354	7274

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 성능 평가 (Multimodal Sentiment Analysis)

Methods	MOSI					MOSEI				
	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
MMIM [11]	0.706	0.798	83.18/84.62	83.12/84.46	47.23	0.540	0.760	82.46/84.98	82.79/84.89	53.44
HyCon [12]	0.709	0.792	82.40/83.20	81.70/83.10	47.80	0.593	0.768	82.16/85.42	81.96/85.02	46.72
PS-Mixer [22]	0.802	0.757	82.51/83.99	82.53/84.07	41.40	0.541	0.766	80.45/85.27	81.13/85.35	52.62
CENet (B) [24]	0.720	0.714	82.36/84.00	82.38/84.06	43.88	0.546	0.765	82.61/85.14	82.83/85.05	52.72
ALMT [23]	0.754	0.774	80.90/83.08	80.75/83.01	44.17	0.539	0.767	83.09/84.51	83.09/84.17	52.20
MISA [15] †						0.549	0.758	80.30/84.84	80.96/84.90	52.80
FDMER [9] †						0.539	0.769	81.00/85.06	81.59/85.09	52.16
MFSA [31] †						0.577	0.741	80.02/82.10	80.01/81.70	52.70
TMT [32] †	0.770	0.756	80.76/82.62	80.70/82.63	44.17	0.567	0.746	81.73/84.53	82.19/84.52	50.85
DMD [16] †	0.720	0.793	83.22/84.30	83.13/84.21	46.36	0.536	0.769	82.46/85.50	82.88/85.04	52.35
TriDiRA	**0.674**	**0.810**	**83.67/85.52**	**83.43/85.34**	**48.98**	**0.529**	**0.775**	**84.85/85.77**	**84.80/85.47**	**53.49**
$\Delta_{second-best}$	↓ 0.032	↑ 0.012	↑ 0.45/0.90	↑ 0.30/0.88	↑ 1.18	↓ 0.007	↑ 0.006	↑ 0.96/0.27	↑ 1.92/0.38	↑ 0.05

고유 특징 안에 있던 비효과적인 특징을 효과적으로 제거

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ 성능 평가 (Multimodal Humor Detection, Multimodal Emotion Classification)

	UR-FUNNY		MELD	
	ACC-2(↑)	F1-Score(↑)	ACC-6(↑)	F1-Score(↑)
UniMSE [26]	–	–	65.09	65.51
MISA [15] †	70.61	–	–	–
FDMER [9] †	71.87	–	–	–
HyCon [12]	68.20	68.05	61.24	60.72
CENet (B) [24]	69.72	69.70	61.26	58.38
ALMT [23]	70.30	70.26	59.89	55.74
MISA [15] †	70.06	69.96	★	★
FDMER [9] †	71.09	71.02	63.79	61.59
MFSA [31] †	70.30	70.24	61.23	60.92
TMT [32]	68.72	68.60	★	★
DMD [16] †	70.01	69.92	★	★
TriDiRA	**72.58**	**72.45**	**65.56**	**63.44**
$\Delta_{second-best}$	↑ 1.49	↑ 1.43	↑ 1.77	↑ 1.85

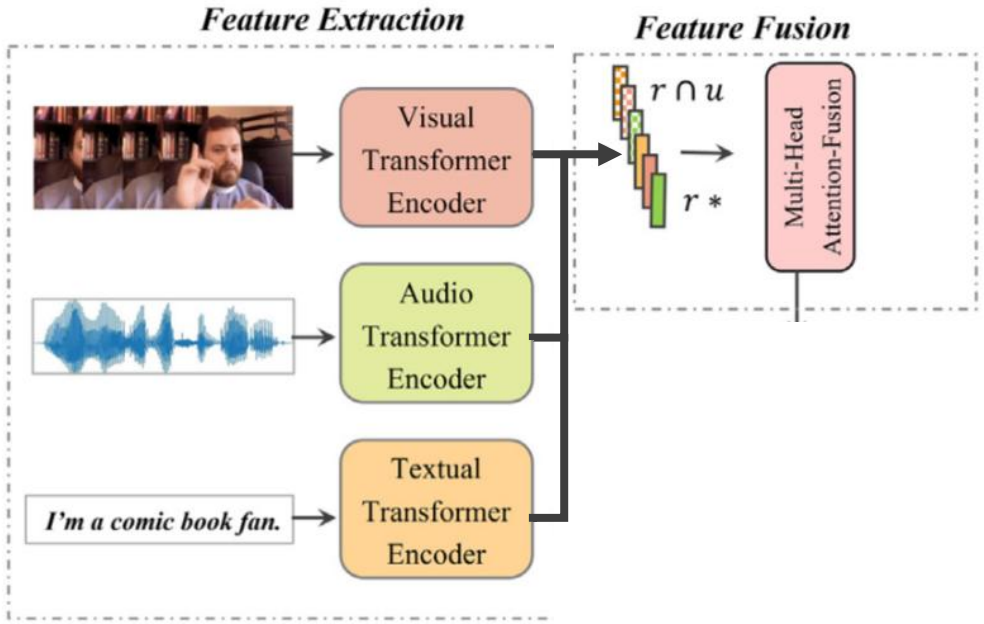
Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Ablation Study

Model	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Importance of Modules					
Baseline	0.711	0.798	81.92/84.30	81.72/84.19	46.50
+ST	0.704	0.803	82.07/84.45	81.76/84.25	46.21
+DS	0.688	0.804	83.07/84.47	83.39/85.29	47.08
Importance of Modalities					
w/o t	1.431	0.041	44.75/42.23	27.67/25.27	15.45
w/o a	0.693	0.801	82.36/84.30	82.17/84.18	46.65
w/o v	0.698	0.800	81.78/83.84	81.62/83.76	48.83
Importance of Regularizations					
w/o $\mathcal{L}_{sim}$	0.686	0.808	82.94/84.21	82.76/83.29	47.81
w/o $\mathcal{L}_{ucorr}$	0.688	0.810	82.65/84.10	82.47/84.02	46.21
w/o $\mathcal{L}_{recon}$	0.694	0.803	81.65/82.60	81.54/82.56	45.63
w/o $\mathcal{L}_{modality}$	0.680	0.807	82.65/84.67	82.47/84.64	48.25
w/o $\mathcal{L}_{h}$	0.680	0.809	82.51/84.62	82.25/84.58	48.40
TriDiRA	**0.674**	**0.810**	**83.67/85.52**	**83.43/85.34**	**48.98**



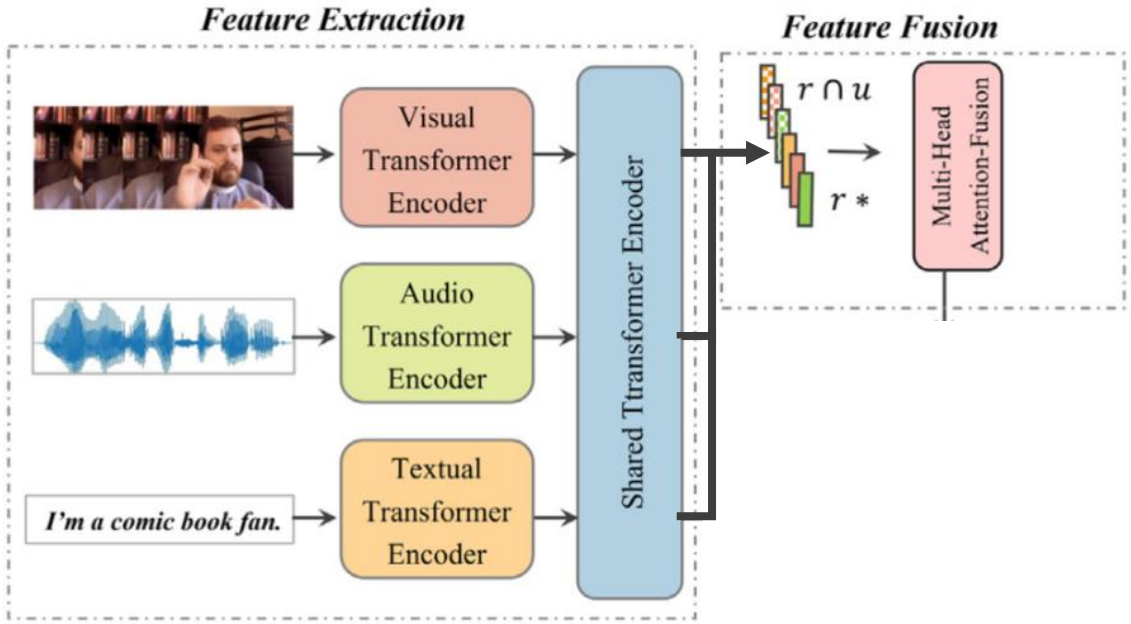
Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Ablation Study

Model	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Importance of Modules					
Baseline	0.711	0.798	81.92/84.30	81.72/84.19	46.50
+ST	0.704	0.803	82.07/84.45	81.76/84.25	46.21
+DS	0.688	0.804	83.07/84.47	83.39/85.29	47.08
Importance of Modalities					
w/o t	1.431	0.041	44.75/42.23	27.67/25.27	15.45
w/o a	0.693	0.801	82.36/84.30	82.17/84.18	46.65
w/o v	0.698	0.800	81.78/83.84	81.62/83.76	48.83
Importance of Regularizations					
w/o $\mathcal{L}_{sim}$	0.686	0.808	82.94/84.21	82.76/83.29	47.81
w/o $\mathcal{L}_{ucorr}$	0.688	0.810	82.65/84.10	82.47/84.02	46.21
w/o $\mathcal{L}_{recon}$	0.694	0.803	81.65/82.60	81.54/82.56	45.63
w/o $\mathcal{L}_{modality}$	0.680	0.807	82.65/84.67	82.47/84.64	48.25
w/o $\mathcal{L}_{h}$	0.680	0.809	82.51/84.62	82.25/84.58	48.40
TriDiRA	**0.674**	**0.810**	**83.67/85.52**	**83.43/85.34**	**48.98**



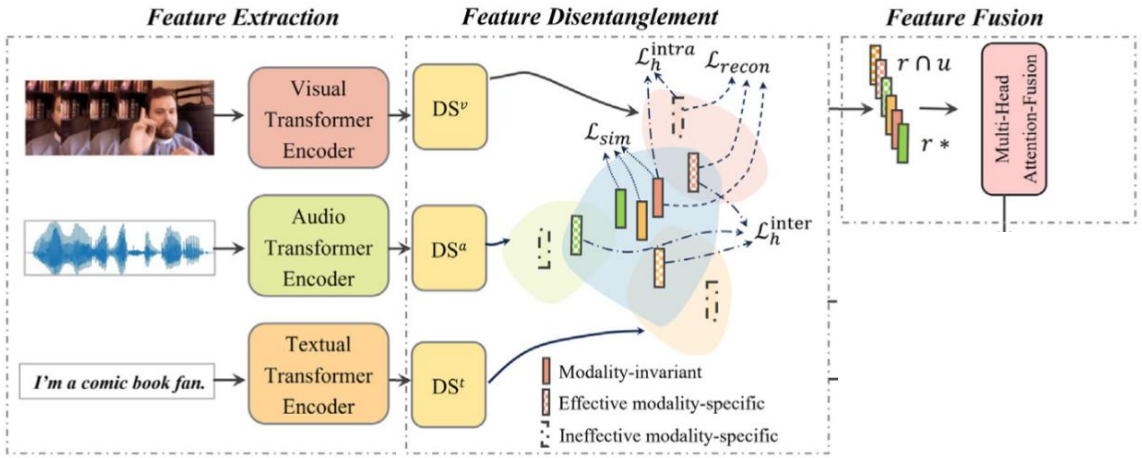
Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Ablation Study

Model	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Importance of Modules					
Baseline	0.711	0.798	81.92/84.30	81.72/84.19	46.50
+ST	0.704	0.803	82.07/84.45	81.76/84.25	46.21
+DS	0.688	0.804	83.07/84.47	83.39/85.29	47.08
Importance of Modalities					
w/o t	1.431	0.041	44.75/42.23	27.67/25.27	15.45
w/o a	0.693	0.801	82.36/84.30	82.17/84.18	46.65
w/o v	0.698	0.800	81.78/83.84	81.62/83.76	48.83
Importance of Regularizations					
w/o $\mathcal{L}_{sim}$	0.686	0.808	82.94/84.21	82.76/83.29	47.81
w/o $\mathcal{L}_{ucorr}$	0.688	0.810	82.65/84.10	82.47/84.02	46.21
w/o $\mathcal{L}_{recon}$	0.694	0.803	81.65/82.60	81.54/82.56	45.63
w/o $\mathcal{L}_{modality}$	0.680	0.807	82.65/84.67	82.47/84.64	48.25
w/o $\mathcal{L}_{h}$	0.680	0.809	82.51/84.62	82.25/84.58	48.40
TriDiRA	**0.674**	**0.810**	**83.67/85.52**	**83.43/85.34**	**48.98**



Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Ablation Study

Model	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Importance of Modules					
Baseline	0.711	0.798	81.92/84.30	81.72/84.19	46.50
+ST	0.704	0.803	82.07/84.45	81.76/84.25	46.21
+DS	0.688	0.804	83.07/84.47	83.39/85.29	47.08
Importance of Modalities					
w/o t	1.431	0.041	44.75/42.23	27.67/25.27	15.45
w/o a	0.693	0.801	82.36/84.30	82.17/84.18	46.65
w/o v	0.698	0.800	81.78/83.84	81.62/83.76	48.83
Importance of Regularizations					
w/o $\mathcal{L}_{sim}$	0.686	0.808	82.94/84.21	82.76/83.29	47.81
w/o $\mathcal{L}_{ucorr}$	0.688	0.810	82.65/84.10	82.47/84.02	46.21
w/o $\mathcal{L}_{recon}$	0.694	0.803	81.65/82.60	81.54/82.56	45.63
w/o $\mathcal{L}_{modality}$	0.680	0.807	82.65/84.67	82.47/84.64	48.25
w/o $\mathcal{L}_{h}$	0.680	0.809	82.51/84.62	82.25/84.58	48.40
TriDiRA	**0.674**	**0.810**	**83.67/85.52**	**83.43/85.34**	**48.98**

Text 모달리티가 다른 모달리티에 비해
여전히 강력한 성능을 유지하고 있음을 시사

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Ablation Study

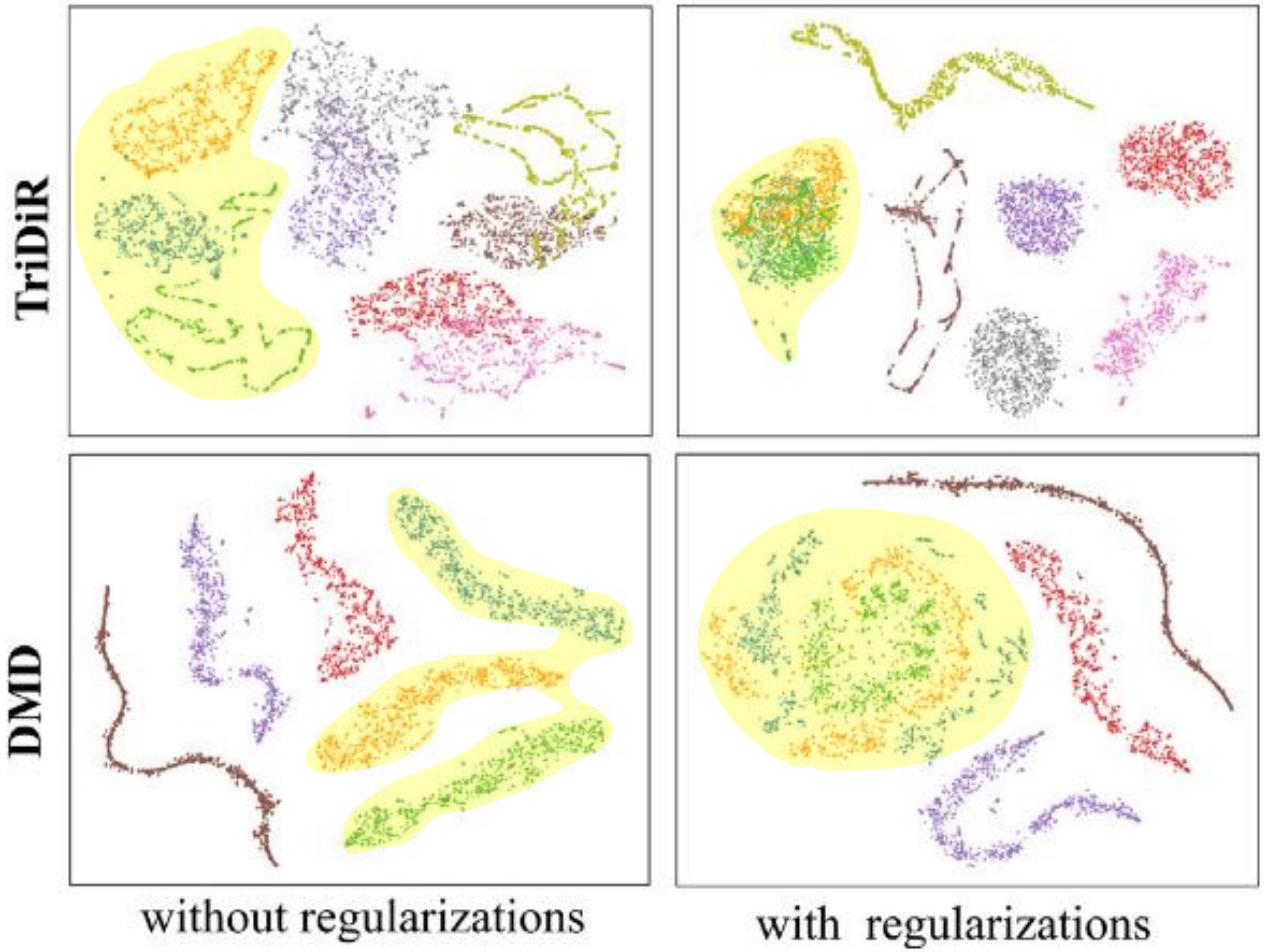
Model	MAE(↓)	Corr(↑)	Acc-2(↑)	F1-Score(↑)	Acc-7(↑)
Importance of Modules					
Baseline	0.711	0.798	81.92/84.30	81.72/84.19	46.50
+ST	0.704	0.803	82.07/84.45	81.76/84.25	46.21
+DS	0.688	0.804	83.07/84.47	83.39/85.29	47.08
Importance of Modalities					
w/o t	1.431	0.041	44.75/42.23	27.67/25.27	15.45
w/o a	0.693	0.801	82.36/84.30	82.17/84.18	46.65
w/o v	0.698	0.800	81.78/83.84	81.62/83.76	48.83
Importance of Regularizations					
w/o $\mathcal{L}_{sim}$	0.686	0.808	82.94/84.21	82.76/83.29	47.81
w/o $\mathcal{L}_{ucorr}$	0.688	0.810	82.65/84.10	82.47/84.02	46.21
w/o $\mathcal{L}_{recon}$	0.694	0.803	81.65/82.60	81.54/82.56	45.63
w/o $\mathcal{L}_{modality}$	0.680	0.807	82.65/84.67	82.47/84.64	48.25
w/o $\mathcal{L}_{h}$	0.680	0.809	82.51/84.62	82.25/84.58	48.40
TriDiRA	**0.674**	**0.810**	**83.67/85.52**	**83.43/85.34**	**48.98**

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

TriDiRA

Triple Disentangled Representation Learning for Multimodal Affective Analysis (2025, Information Fusion)

❖ Subspace visualization (t-SNE)



invariant: v_c a_c t_c

effective specific: v_s a_s t_s

ineffective specific: v_i a_i t_i

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

Summary

❖ How to handle multi-modal heterogeneity and improve performance using disentangled learning?

1. MISA (2020, ACM Multimedia): 멀티모달에 대한 이중 분리 방법론 최초 제안
   - 멀티모달 특징을 공통 특징과 고유 특징 두 가지로 분리하는 이중 분리 개념을 정립
   - 고유 특징 내에 task에 비효과적이거나 상충하는 정보가 포함될 수 있다는 한계

2. DMD (2023, CVPR): 그래프 증류를 통한 분리 표현 고도화
   - 그래프 증류(Graph Distillation)라는 동적 지식 전달 매커니즘을 도입해 강한 모달리티가 약한 모달리티를 가르치도록 제안
   - MISA와 같이 고유 특징 전체가 유용하다고 가정해, 특징에 포함된 비유효한 정보까지 증류에 활용될 수 있는 한계

3. TriDiRA (2025, Information Fusion): 삼중 분리 방법론을 제안
   - 이중 분리로 획득한 고유 특징을 다시 효과적인 부분과 효과적이지 않은 부분으로 나누는 삼중 분리 방법론을 제안
   - Task와 무관한 특징을 제거해 정보의 질을 높이고자 함

Zhou, Y., Liang, X., Chen, H., Zhao, Y., Chen, X., & Yu, L. (2025). Triple disentangled representation learning for multimodal affective analysis. *Information Fusion*, *114*, 102663. https://doi.org/10.1016/j.inffus.2024.102663

고맙습니다