ReconBoost: Boosting Can Achieve Modality Reconcilement

Cong Hua, Qianqian Xu*, Shilong Bao,

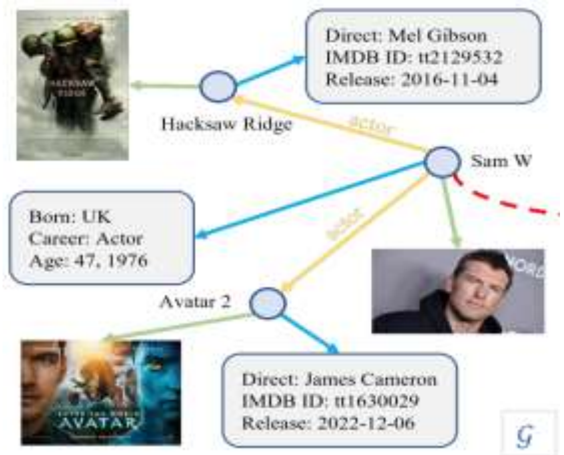
Zhiyong Yang, Qingming Huang*

Cong Hua

2024.07

Background

☐ Real-world data usually follows a multi-modal nature

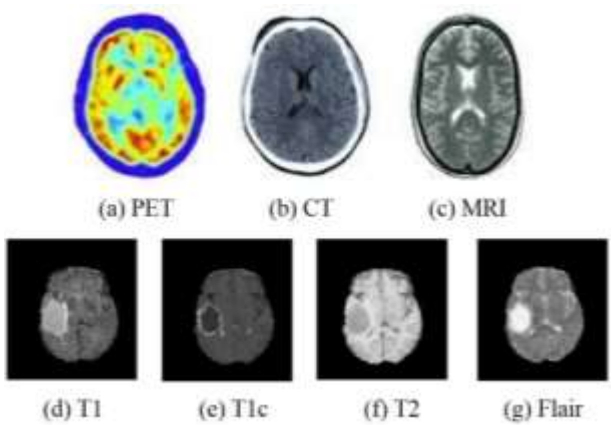


Knowledge Graph

Visual

Audio

Scene Understanding

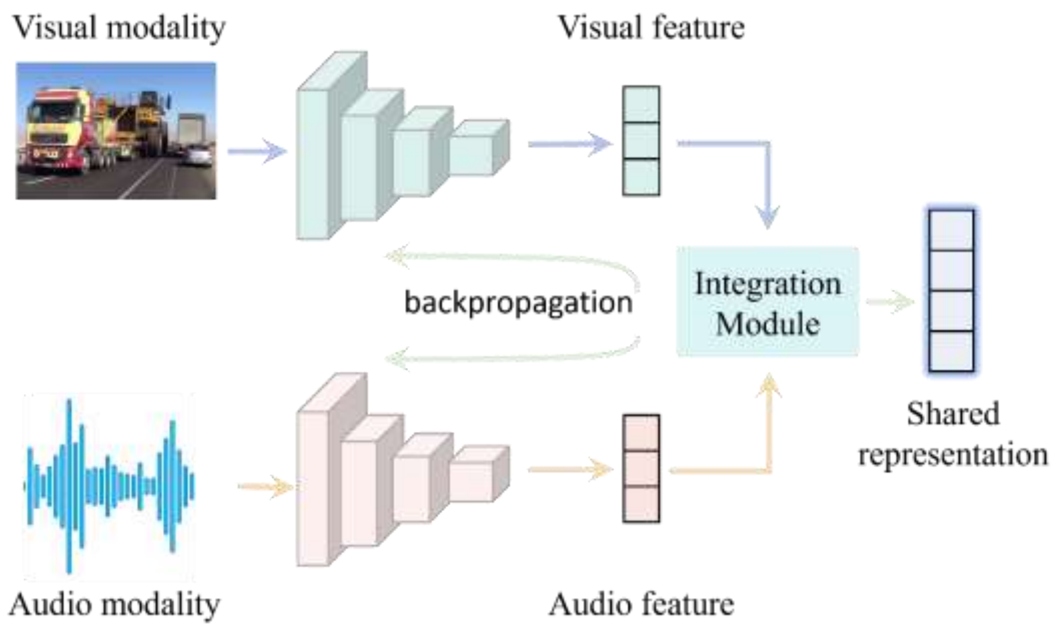


Medical Diagnosis

Cross-modal Retrieval

Background

◻ Multi-modal joint learning



- The prevailing paradigm in multi-modal learning typically employs a **joint learning** strategy.
- Various MML studies focus on integrating modality-specific features into **a shared representation** for downstream tasks.

Background

◻ Modality Competition

The gradient update rule of k-th modality learner

$$\begin{aligned}\theta_k^{t+1} &= \theta_k^t - \eta \cdot \nabla_{\theta_k^t} \mathcal{L}(\Phi_M^t(x), y) \\ &= \theta_k^t - \eta \cdot \frac{1}{N}\sum_{i=1}^{N}\left(\frac{\partial W_k \cdot \mathcal{F}_k(\theta_k^t; m_i^k)}{\partial \theta_k^t}\right)^{\top} \cdot \underbrace{\frac{\partial \ell\left(\Phi_M^t(x_i), y_i\right)}{\partial \Phi_M^t(x_i)}}_{\text{shared}}\end{aligned}$$

dominant modality

If $\frac{\partial \ell(\phi_k^t(x_i), y_i)}{\partial \phi_k^t(x_i)}$ approximates $\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)}$,

this modality will converge fast and overpower the learning process.

weak modality

If $\frac{\partial \ell(\phi_k^t(x_i), y_i)}{\partial \phi_k^t(x_i)}$ does not approximate $\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)}$,

this modality will be stuck at bad local optimums.

Background

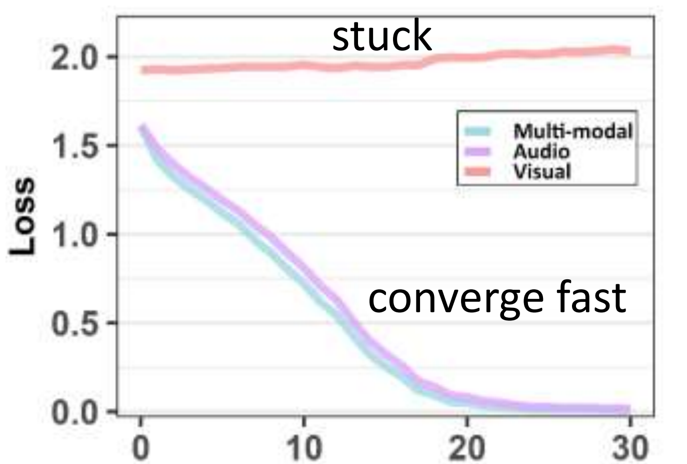
◻ Modality Competition – Empirical Observation

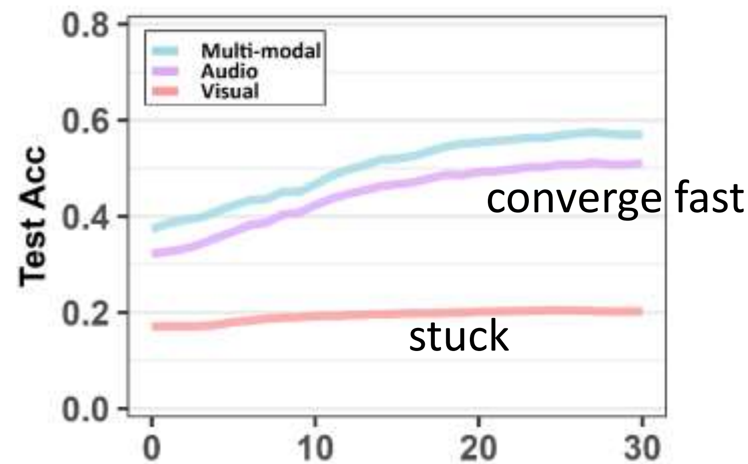
- The gradient of audio $\frac{\partial \ell(\phi_k^t(x_i), y_i)}{\partial \phi_k^t(x_i)}$ approximates $\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)}$

  Audio modality will converge fast and overpower the learning process.

- The gradient of visual $\frac{\partial \ell(\phi_k^t(x_i), y_i)}{\partial \phi_k^t(x_i)}$ does not approximate $\frac{\partial \ell(\Phi_M^t(x_i), y_i)}{\partial \Phi_M^t(x_i)}$

  Visual modality will be stuck at bad local optimums.



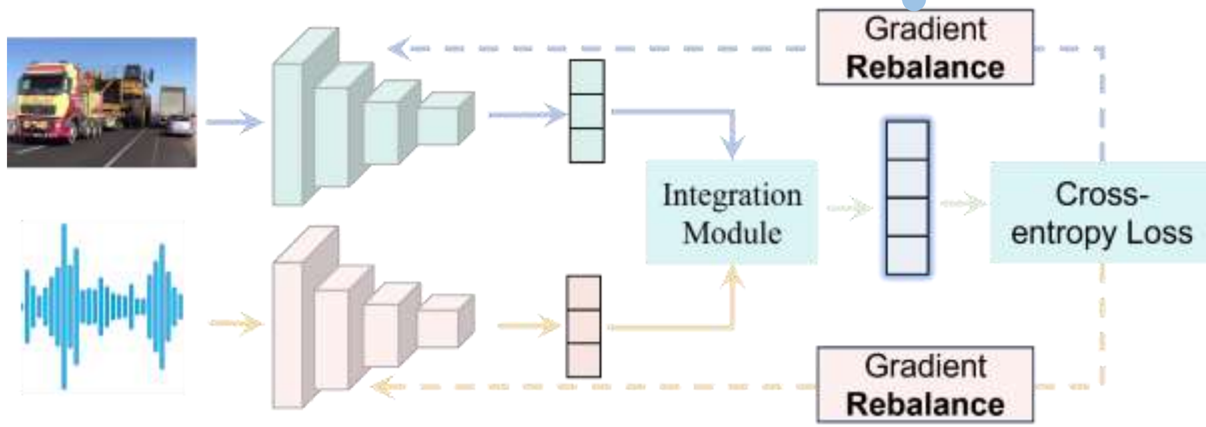


Performance on Audio-Visual dataset

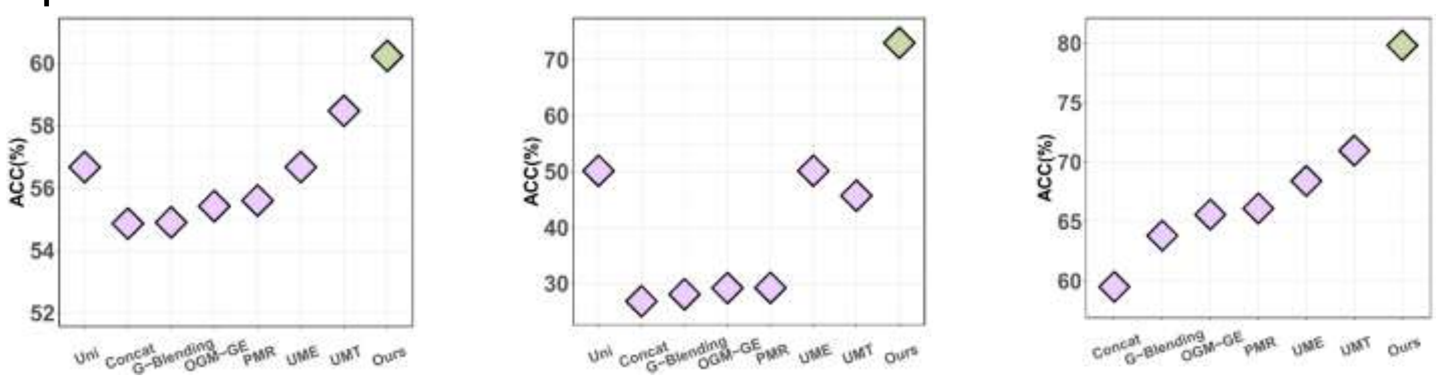
Related works

◻ Balanced multi-modal learning

- The primary concern is how to **balance** optimization progress across multi-modal learners.



- Given the nature of **joint optimization**, only limited improvements can be achieved.

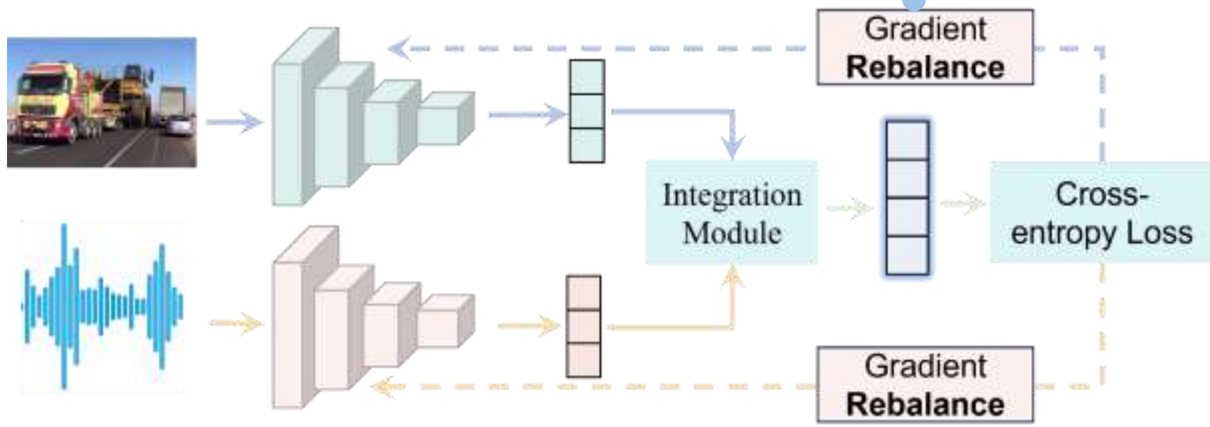


(a) Audio Modality (b) Visual Modality (c) Multi-modal

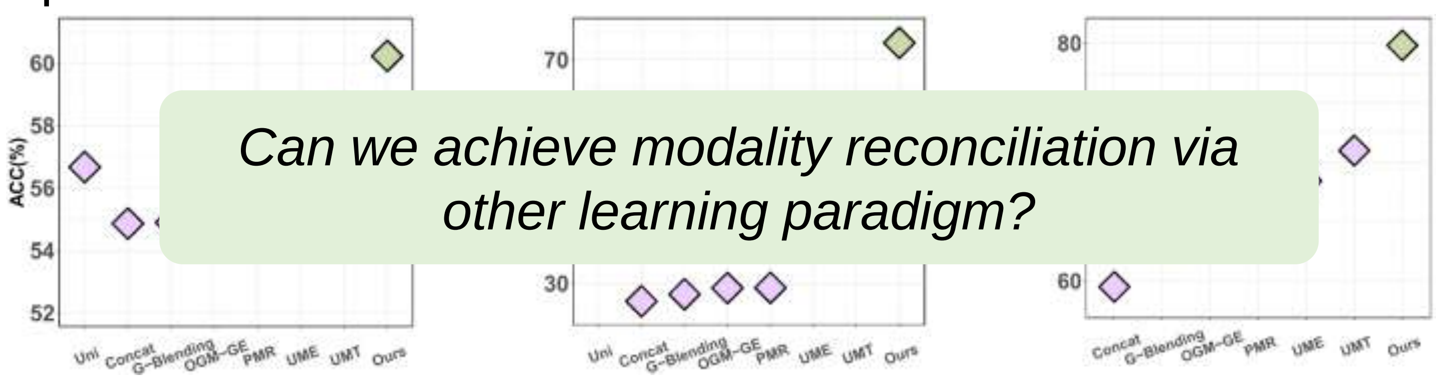
Related works

◻ Balanced multi-modal learning

- The primary concern is how to **balance** optimization progress across multi-modal learners.



- Given the nature of **joint optimization**, only limited improvements can be achieved.



Can we achieve modality reconciliation via other learning paradigm?

(a) Audio Modality (b) Visual Modality (c) Multi-modal

Main part

Naive version of modality-alternating learning

Step 1: Each time, we pick a specific modality learner ϕ_k to **update**, and keep other fixed.

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \mathcal{L}\left(\phi_k^t\left(m^k\right), y\right)$$

$$\mathcal{L}\left(\phi_k^t\left(m^k\right), y\right) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(\phi_k\left(\vartheta_k^t; m_i^k\right), y_i\right)$$

Step 2: Multi-modal scores are **merged** to obtain the final score.

$$\Phi_M\left(x_i\right) = \sum_{k=1}^{M} \phi_k\left(\vartheta_k; m_i^k\right)$$

- The gradient across different modalities are naturally disentangled from each other, alleviating the modality competition issue.
- This approach ensures the exploitation of uni-modal features, but **neglects** the investigation of cross-modal interaction.

Main part

□ Naive version of modality-alternating learning

Step 1: Each time, we pick a specific modality learner ϕ_k to **update**, and keep other fixed.

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \mathcal{L}(\phi_k^t(m^k), y)$$

$$\mathcal{L}(\phi_k^t(m^k), y) = \frac{1}{N} \sum_{i=1}^{N} \ell\left(\phi_k(\vartheta_k^t; m_i^k), y_i\right)$$

Step 2: Multi-modal scores are **merged** to obtain the final score.

$$\Phi_M(x_i) = \sum_{k=1}^{M} \phi_k\left(\vartheta_k; m_i^k\right)$$

- The gradient across different modalities are naturally disentangled fr
- T it neglect the investigation of cross-modal diversity.

How to design a more effective modality supervised signal?

Main part

Modality-alternating Update with Dynamic Reconcilment

Step 1: Each time, we pick a specific modality learner ϕ_k to **update**, and keep other fixed.

$$\vartheta_k^{t+1} = \vartheta_k^t - \eta \cdot \nabla_{\vartheta_k^t} \tilde{\mathcal{L}}_s(\phi_k^t(m^k), y)$$

$$\tilde{\mathcal{L}}_s(\phi_k(m^k), y) = \frac{1}{N}\sum_{i=1}^{N}\left[\underbrace{\ell(\phi_k(\vartheta_k; m_i^k), y_i)}_{\text{agreement term}} - \lambda \cdot \underbrace{\mathbb{D}_s(\Phi_{M/k}(x_i), \phi_k(\vartheta_k; m_i^k))}_{\text{reconcilement regularization term}}\right]$$

Dynamically maintain the trade-off between two items:

- The agreement term aligns the overall predictor with the ground truth.
- The reconcilement regularization term investigates the cross-modal diversity.

Step 2: Multi-modal scores are **merged** to produce the final score.

$$\Phi_M(x_i) = \sum_{k=1}^{M} \phi_k(\vartheta_k; m_i^k)$$

Main part

Connection to the Boosting Strategy

The overall optimization property of ReconBoost is **unclear**

$$\tilde{\mathcal{L}}_s(\phi_k(m^k),y) = \frac{1}{N}\sum_{i=1}^{N}\left[\underbrace{\ell(\phi_k(\vartheta_k;m_i^k),y_i)}_{\text{agreement term}} - \lambda\cdot\underbrace{\mathbb{D}_s(\Phi_{M/k}(x_i),\phi_k(\vartheta_k;m_i^k))}_{\text{reconcilement regularization term}}\right]$$

Theorem 1. Connection to the Gradient Boosting (GB) method

Let the reconcilement regularization be a KL divergence function:

$$\mathbb{D}_s(\Phi_{M/k}(x_i),\phi_k(\vartheta_k;m_i^k)) = \mathbb{D}_{KL,s}(\Phi_{M/k}(x_i)|\phi_k(\vartheta_k;m_i^k))$$

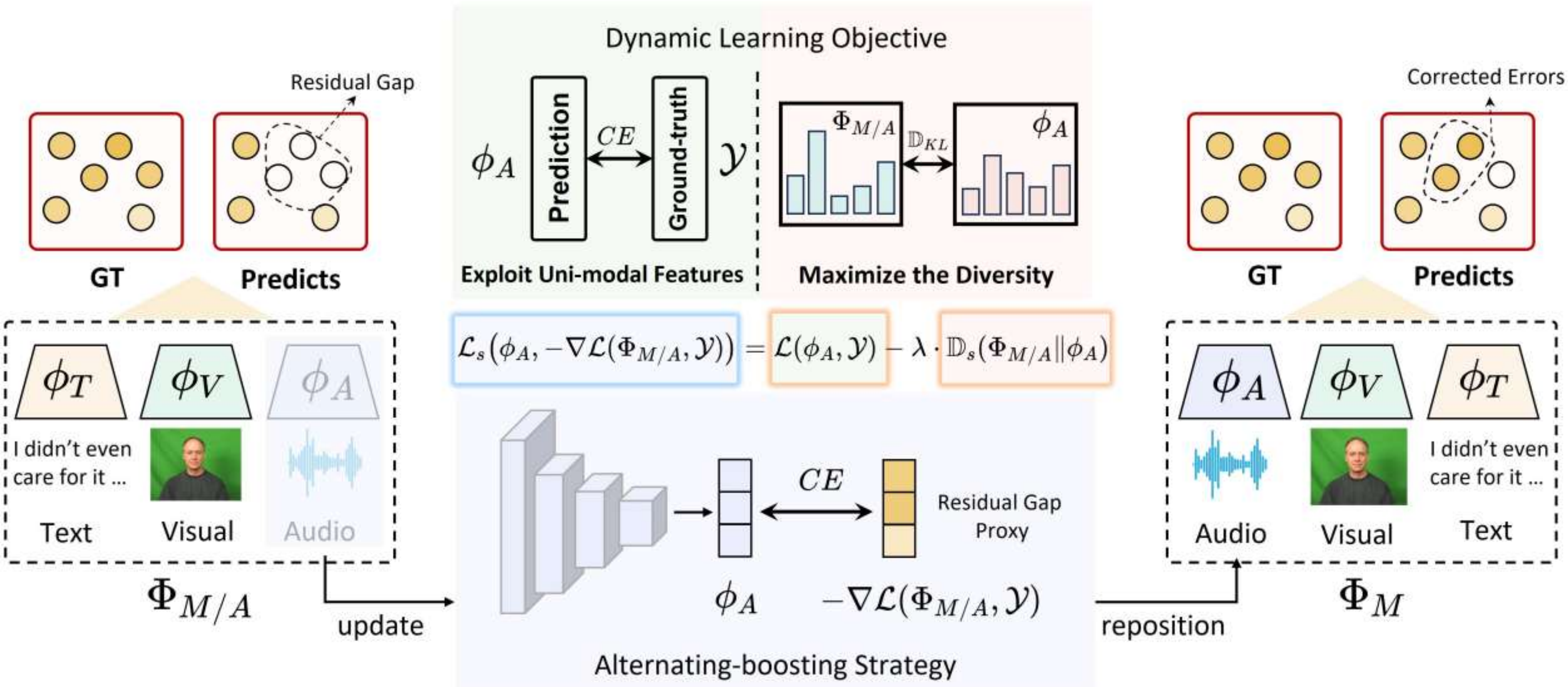
Then,

$$\nabla_{\vartheta_k}\tilde{\mathcal{L}}_s(\phi_k(m^k),y) \Longleftrightarrow \nabla_{\vartheta_k}\mathcal{L}(\phi_k(m^k), -\nabla_{\Phi_{M/k}}\ell(\Phi_{M/k}(x),y))$$

- Optimizing the dynamic loss functions $\tilde{\mathcal{L}}$ in ReconBoost **consistently** optimizes the original loss $\mathcal{L}$ with a progressively changing pseudo-label in GB algorithm (Friedman, 2001).
- The updated modality learner can focus on the errors made by others.
- ReonBoost only preserves the last learner of each modality, formulating **alternating-boosting strategy.**

Main part

◻ Pipeline of ReconBoost



- ReconBoost can realize an alternating version of the well-known gradient boosting algorithm.
- ReconBoost purses a reconciliation between the exploitation of uni-modal features and the exploration of cross-modal diversity.

Experiments

Quantitative Comparisons

Method	AVE	CREMA-D	MN40	MOSEI	MOSI	CH-SIMS
AudioNet	59.37	56.67		52.29	54.81	58.20
VisualNet	30.46	50.14	80.51	50.35	57.87	63.02
TextNet	–	–	–	66.41	75.94	70.45
Concat Fusion	62.68	59.50	83.18	66.71	76.23	71.55
G-Blending	62.75	63.81	84.56	66.93	76.45	71.55
OGM-GE	62.93	65.59	85.61	66.67	76.01	71.10
PMR	64.20	66.10	86.20	66.41	76.12	70.90
UME	66.92	68.41	85.37	63.88	76.97	71.77
UMT	67.71	70.97	90.07	67.04	75.80	71.55
Ours	**71.35**	**79.82**	**91.78**	**68.61**	**77.96**	**73.88**

Experiments

☐ Modality-specific Encoder Evaluation

Method	CREMAD Dataset				AVE Dataset			
	Audio	Visual	MIR	DMC	Audio	Visual	MIR	DMC
Uni-train	56.67	50.14	1.13	-	59.37	30.46	1.95	-
Concat Fusion	54.86	26.81	2.05	1.81	55.47	23.96	2.32	1.19
G-Blending	54.90	28.05	1.96	1.73	55.80	24.12	2.31	1.19
OGM-GE	55.42	29.17	1.90	1.68	56.51	25.52	2.21	1.14
PMR	55.60	29.21	1.90	1.68	57.20	26.30	2.17	1.12
UMT	58.47	45.69	1.28	1.13	60.70	31.07	1.95	1.00
Ours	**60.23**	**73.01**	**0.82**	**0.73**	**61.20**	**39.06**	**1.57**	**0.80**

Conclusion

- **Methodologically:** propose a novel multi-modal alternating learning paradigm to address notorious modality competition issue.

- **Theoretically:** show that by choosing a KL-divergence-based reconcilement term, our proposed method can realize an alternating version of the well-known gradient boosting method.

- **Empirically:** Comprehensive experiments justify the effectiveness of our proposed framework on various multi-modal scenarios.

Thanks for your listening!

code