IBMA: Information Bottleneck-Based Multimodal Alignment

Yancheng Wang[†], Zeyu Dong[†], Dongfang Sun[†], Alvin C. Silva[‡], Teresa Wu[†], Yingzhen Yang[†]

[†]Arizona State University [‡]Mayo Clinic Arizona

Motivation: Multimodal Representation Learning

Multimodal learning integrates heterogeneous data sources, e.g., images, text, audio, video, histology, and gene expression.

Key challenge:

- Different modalities provide complementary semantic information.
- Each modality also contains noise, redundancy, and modality-specific artifacts.
- Effective models should keep task-relevant shared information and suppress irrelevant modality-specific variation.

Information Bottleneck Principle

Learn a representation Z that is:

predictive of Y but compressed from X.

A standard IB objective reduces

$$I(Z, X) - I(Z, Y).$$

Multimodal Goal

Learn aligned modality-specific features and a discriminative fused representation.

Limitations of Existing Multimodal IB Methods

Existing methods such as MIB, MCIB, and OMIB mainly apply IB to the fused multimodal representation.

Limitations:

- They do not explicitly align each modality-specific encoder.
- Modality-specific noise can remain in individual representations.
- VAE-style or Gaussian latent assumptions may be restrictive for complex DNN features.

Motivation for IBMA:

- Apply IB not only to the fused feature.
- Also align modality-specific representations using cross-modal supervision.

Grad-CAM example on MELD: IBMA attends more to emotion-relevant regions than OMIB.

IBMA: Main Idea and Notation

Training data:

$$\{X_i^{(1)}, X_i^{(2)}, Y_i\}_{i=1}^n,$$

where $X_i^{(j)}$ is the input from modality j, and Y_i is the label.

Modality-specific representations:

$$Z_i^{(j)} = f^{(j)}(X_i^{(j)}), \qquad j \in \{1, 2\}.$$

Fused multimodal representation:

$$Z_i = F(Z_i^{(1)}, Z_i^{(2)}),$$

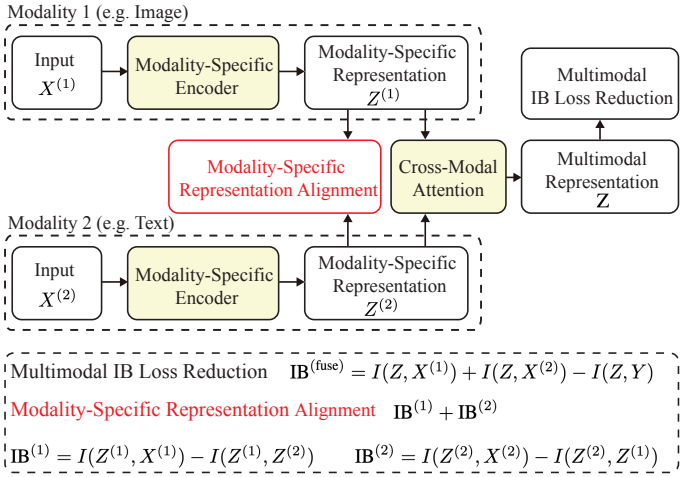
where $F(\cdot, \cdot)$ is a cross-modal attention fusion module.

IBMA applies IB at two levels

$$\mathrm{IB}^{(\mathrm{fuse})} = I(Z, X^{(1)}) + I(Z, X^{(2)}) - I(Z, Y),$$

$$\mathrm{IB}^{(j)} = I(Z^{(j)}, X^{(j)}) - I(Z^{(j)}, Z^{(j')}), \qquad j' \neq j.$$

IBMA Framework



- Each modality is encoded by a modality-specific encoder.
- Cross-modal attention produces the fused multimodal representation.
- IBMA regularizes both the fused representation and each modality-specific representation.

Prototype-Based Feature Distribution Modeling

IBMA estimates feature distributions using learnable prototypes rather than imposing a Gaussian prior. For modality j, let $\{\mathcal{F}_a^{(j)}\}_{a=1}^{C}$ be C learnable prototypes. The soft assignment of $Z_i^{(j)}$ to prototype a is:

$$\phi(Z_i^{(j)}, a) = \frac{\exp\left(-\|Z_i^{(j)} - \mathcal{F}_a^{(j)}\|_2^2\right)}{\sum_{a'=1}^{C} \exp\left(-\|Z_i^{(j)} - \mathcal{F}_{a'}^{(j)}\|_2^2\right)}.$$

This induces empirical probabilities such as:

$$\mathbb{P}(Z^{(j)} \in a) = \frac{1}{n} \sum_{i=1}^{n} \phi(Z_i^{(j)}, a),$$

$$\mathbb{P}(Z^{(j)} \in a, X^{(j)} \in y), \qquad \mathbb{P}(Z^{(j)} \in a, Z^{(j')} \in y).$$

Why prototypes?

They provide a distribution-free way to estimate mutual information over learned representations.

IBB: Distribution-Free Upper Bound for IB Loss

For modality j, IBMA derives:

$$\mathrm{IB}^{(j)} \leq \mathrm{IBB}^{(j)} = \frac{1}{n} \sum_{i=1}^{n} \left(U_i^{(j)} - V_i^{(j)} \right).$$

$$U_i^{(j)} = \sum_{a=1}^{C} \sum_{y=1}^{C} \phi_j(i, a, y) \log \left(\frac{\phi_j(i, a, y)}{p_y \phi(Z_i^{(j)}, a)} \right),$$

$$V_i^{(j)} = \sum_{a=1}^{C} \sum_{y=1}^{C} \phi_{jj'}(i, a, y) \log Q^{(j)} \left(Z^{(j)} \in a \mid Z^{(j')} \in y \right).$$

- $p_y = \mathbb{P}(X^{(j)} \in y)$ is the empirical class/prototype probability.
- $\phi_j(i, a, y) = \phi(Z_i^{(j)} \in a, X_i^{(j)} \in y)$.
- $\phi_{jj'}(i, a, y) = \phi(Z_i^{(j)} \in a, Z_i^{(j')} \in y)$.
- $Q^{(j)}(Z^{(j)} \in a \mid Z^{(j')} \in y)$ is a variational conditional distribution.

Training Objective and Efficiency

For mini-batch $\mathcal{B}_b$, IBMA optimizes:

$$\mathcal{L}_b = \mathrm{CE}_b + \eta \mathrm{IBB}_b,$$

where

$$\mathrm{IBB}_b = \mathrm{IBB}_b^{(1)} + \mathrm{IBB}_b^{(2)} + \mathrm{IBB}_b^{(\mathrm{fuse})},$$

and

$$\mathrm{CE}_b = \mathrm{CE}_b^{(1)} + \mathrm{CE}_b^{(2)} + \mathrm{CE}_b^{(\mathrm{fuse})}.$$

What the CE term does

- Maintains task discriminability.
- Supervises modality-specific and fused predictions.

What the IBB term does

- Suppresses redundant input information.
- Aligns modality-specific representations.
- Improves fused multimodal features.

Computational advantage

IBB has complexity $\Theta(nT_0 + nC^2)$, while CLUB-style distribution-free estimation requires $\Theta(n^2 T_0)$.

Experimental Tasks, Datasets, and Summary Results

Evaluated multimodal settings:

- **Emotion recognition:** CREMA-D, MELD, IEMOCAP.
- **Disease classification:** MIMIC-CXR, CheXpert.
- **Sentiment analysis:** CMU-MOSI, MELD.
- **Anomalous tissue detection:** 10x-hBC-A–D.
- **Four-modality classification:** PME4 with audio, video, EEG, and EMG.
- **Large-class multimodal classification:** UPMC Food-101 and WIKI-DOC.

Task	Dataset	Metric	Best Baseline	**IBMA**
Emotion Recognition	CREMA-D	Accuracy	63.6	**65.4 ± 0.4**
Emotion Recognition	MELD	Accuracy	64.3	**66.3 ± 0.3**
Emotion Recognition	IEMOCAP	Accuracy	74.3	**75.7 ± 0.3**
Disease Classification	MIMIC-CXR	mAUC	71.0	**72.7 ± 0.3**
Disease Classification	CheXpert	mAUC	89.3	**91.1 ± 0.2**
Sentiment Analysis	CMU-MOSI	Acc-2 / F1	86.9 / 87.2	**87.9 / 88.3**
Sentiment Analysis	MELD	Acc-2 / F1	80.5 / 80.1	**82.0 / 81.5**
Anomalous Tissue Detection	10x-hBC Mean	AUC / F1	0.754 / 0.737	**0.774 / 0.754**
Four-Modality Classification	PME4	Accuracy	81.1	**82.2**
Large-Class Classification	UPMC / WIKI-DOC	Accuracy	91.2 / 93.0	**92.4 / 94.1**

Key Ablation and Takeaway

Ablation: modality-specific alignment

- MIMIC-CXR:

$$72.7 \rightarrow 71.4$$

without modality-specific representation alignment.

- CheXpert:

$$91.1 \rightarrow 89.7$$

without modality-specific representation alignment.

Conclusion: Cross-modal guidance improves individual encoders, which further improves fused predictions.

Ablation: IBB upper bound

- CREMA-D accuracy:

$$\text{VIB} : 63.9, \quad \text{APIB} : 63.8, \quad \text{CLUB} : 64.1,$$

$$\textbf{IBB} : \textbf{65.4}.$$

- IBB achieves stronger accuracy while using only 52.6% of CLUB training time.

Takeaway

IBMA improves multimodal learning by combining modality-specific IB alignment with an efficient distribution-free IB upper bound.

Conclusion

- IBMA introduces modality-specific IB alignment to guide each encoder with complementary cross-modal information.
- The proposed IBB provides an efficient, distribution-free upper bound for optimizing IB loss without Gaussian latent assumptions.
- Experiments across diverse multimodal tasks show consistent improvements over strong fusion, contrastive, and IB-based baselines.

IBMA: aligned modality-specific features for stronger multimodal prediction.