InfMasking: Unleashing Synergistic Information by Contrastive Multimodal Interactions

Liangjian Wen[1,2], Qun Dai[1], Jianzhuang Liu[3], Jiangtao Zheng[1], Yong Dai[4], Dongkai Wang[1], Zhao Kang[5], Jun Wang[1], Zenglin Xu[6,7], Jiang Duan[1]

[1] Southwestern University of Finance and Economics
[2] Engineering Research Center of Intelligent Finance, Ministry of Education
[3] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences [4] X-Humanoid
[5] University of Electronic Science and Technology of China [6] Shanghai Academy of AI for Science
[7] Artificial Intelligence Innovation and Incubation Institute, Fudan University

Outline

- **Background**
- **Method**
- **Experimental Results**
- **Conclusion**

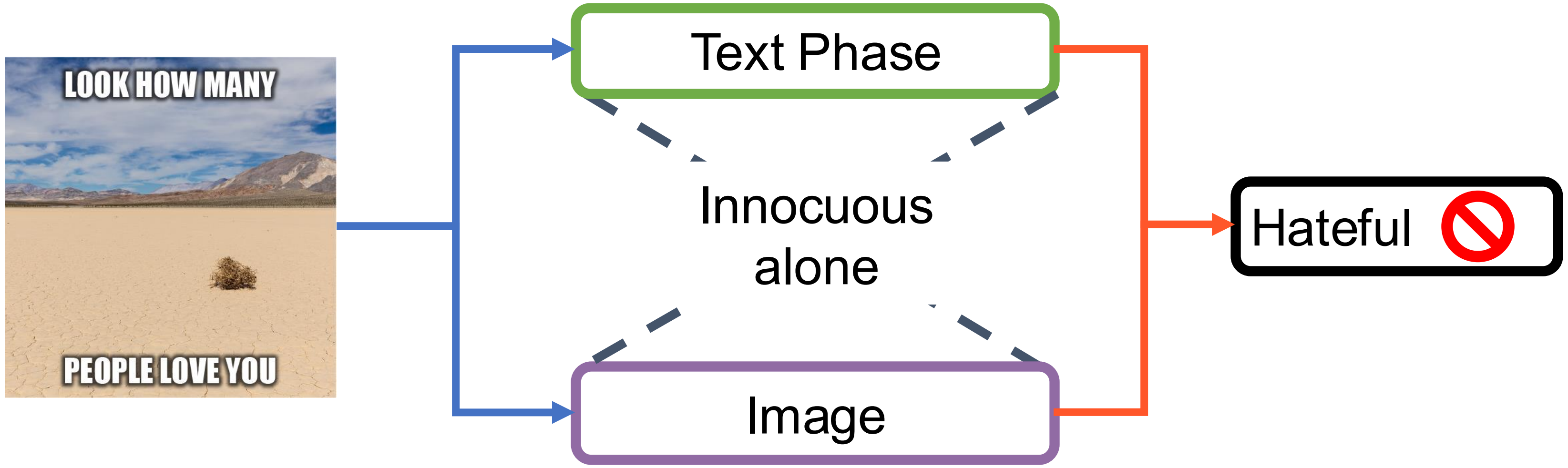
Outline

- **Background**
- **Method**
- **Experimental Results**
- **Conclusion**

Multimodal Interactions

Three fundamental Interactions：

- **R**edundancy: modality independence via shared info
- **U**niqueness: exclusive modality info for task
- **S**ynergy: combined modalities for complementary outcome



Multimodal Interactions Problems

- Existing multimodal contrastive learning methods mostly rely on the following assumption:

Definition 1 (***Multi-view redundancy***) $\exists\, \varepsilon > 0$ *such that* $I(Y; X_1|X_2) < \varepsilon$ *and* $I(Y; X_2|X_1) < \varepsilon$.

This assumption only enables the model to learn the redundant information R.

- Recent works have attempted to learn full multimodal interactions, yet they primarily emphasize enhanced redundant and unique interactions (R & U).

Outline

- Background
- **Method**
- Experimental Results
- Conclusion

Preliminaries

- Consider two modalities X_1 and X_2 and a task Y
- According to PID, the mutual information $I(X_1, X_2; Y)$ can be decomposed as:

$$I(X_1, X_2; Y) = R + S + U_1 + U_2,$$

  where R represents redundant information, S represents synergistic information and U1 and U2 represent unique information specific to X1 and X2, respectively

- This decomposition is supported by consistency equations derived from the chain rule of mutual information:

$$I(X_1; Y) = R + U_1,\ \ I(X_2; Y) = R + U_2,\ \ I(X_1; X_2; Y) = R - S,$$

Preliminaries

- In self-supervised learning, Y remains unspecified, , presenting a unique challenge
- Multimodal Redundancy Assumption:

Assumption 1 (***Minimal label-preserving multimodal augmentations***) *A set $\mathcal{T}^*$ of multimodal transformations exists, such that for any $t \in \mathcal{T}^*$ and $X' = t(X)$, the mutual information $I(X; X') = I(X; Y)$ holds, preserving the information with label Y.*

- Defining a multimodal latent variable $Z_\theta = f_\theta(X)$ and $Z'_\theta = f_\theta(X')$.
- Considering the Markov chains: $X \to X' \to Z'_\theta$ and $Z'_\theta \to X \to Z_\theta$, we can establish the following mutual information bounds:

$$I(Z_\theta; Z'_\theta) \leq I(X; Z'_\theta) \leq I(X; X').$$

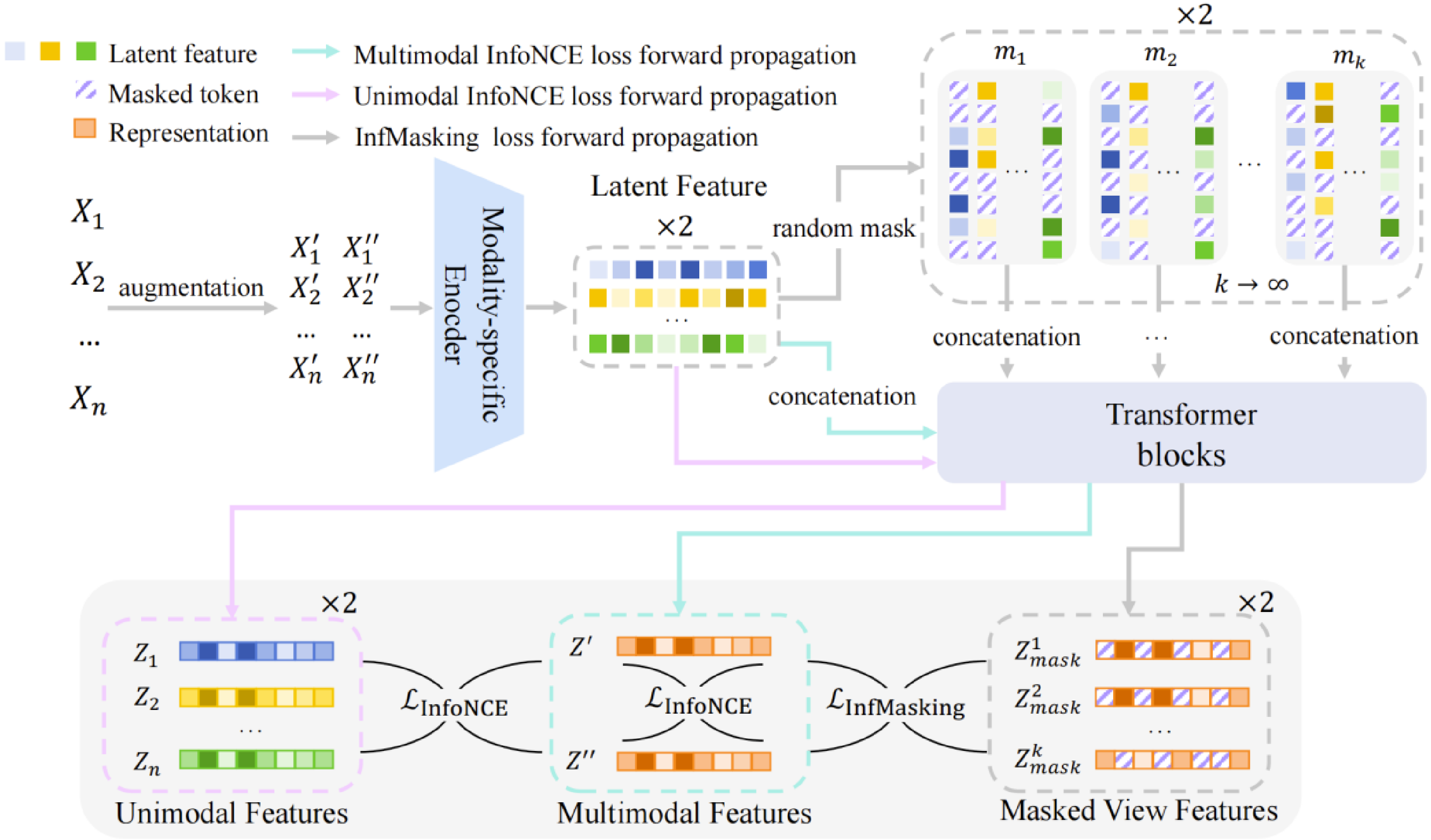
Preliminaries

- According to these inequalities and Assumption 1, we can prove the following lemmas

Lemma 1 *When optimizing the function f_θ to maximize mutual information $I\left(Z_\theta; Z'_\theta\right)$, and under the assumption that the network f_θ possesses sufficient expressivity, we observe that in the optimal parameter configuration:* $I(Z_{\theta^\star}, Z'_{\theta^\star}) = I(X, X') = I(X, Y)$.

Lemma 2 *Suppose f_{θ^*} is optimal, meaning it maximizes $I(Z_{\theta^\star}, Z'_{\theta^\star})$. Then, the equality $I(Z_{\theta^\star}, Y) = I(X', Y)$ holds. For the special case where $T = \{t_i\}$ such that $X' = t_i(X) = X_i$ and $Z'_{\theta^\star} = f_{\theta^\star}(X) = Z_i$ for $i \in \{1, 2\}$, the following holds:$I(Z_i; Y) = I(X_i; Y) = R + U_i$.*

Overview of InfMasking framework



$$\mathcal{L}_{\text{Total loss}} = -\underbrace{\hat{I}_{\text{NCE}}(Z', Z'')}_{\approx R+S+\sum_{i=1}^{n} U_i} - \sum_{i=1}^{n} \underbrace{\frac{1}{2}\left(\hat{I}_{\text{NCE}}(Z_i, Z') + \hat{I}_{\text{NCE}}(Z_i, Z'')\right)}_{\approx R+U_i} \underbrace{-\mathbb{E}_{\text{mask}}\left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''_{\text{mask}}, Z'')\right]}_{\mathcal{L}_{\text{InfMasking}}}$$

Contrastive Synergistic Information via Infinite Masking

- During the fusion process, we continuously and randomly mask a significant portion of the features from each modality an infinite number of times to capture synergistic information.

$$
\begin{aligned}
\mathcal{L}_{\text{InfMasking}} &= -\lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \hat{I}_{\text{NCE}}(Z'^{\,k}_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''^{\,k}_{\text{mask}}, Z'') \\
&= -\mathbb{E}_{\text{mask}} \left[\hat{I}_{\text{NCE}}(Z'_{\text{mask}}, Z') + \hat{I}_{\text{NCE}}(Z''_{\text{mask}}, Z'') \right].
\end{aligned}
$$

computationally expensive !

Lemma 3 *Let $\boldsymbol{\mu}_{z'_{mask}}$ and $\boldsymbol{\Sigma}_{z'_{mask}}$ be the Gaussian mean vector and covariance matrix of z'_{mask}, respectively. The lower bound of $\mathbb{E}_{mask}\left[\hat{I}_{NCE}(Z'_{mask}, Z')\right]$ can be obtained as follows:* $\mathbb{E}_{mask}[\hat{I}_{NCE}(Z'_{mask}, Z')]$

$$
\geq \mathbb{E}_{z'\sim p(Z')} \left[z'^T \boldsymbol{\mu}_{z'_{mask}}/\tau - \log[\exp(z'^T \boldsymbol{\mu}_{z'_{mask}}/\tau + \frac{z'^T \boldsymbol{\Sigma}_{z'_{mask}} z}{2\tau^2}) + \sum_{z'_{neg}} \exp(z'^T_{neg} z'_{mask}/\tau)] \right]
$$

Outline

- Background
- Method
- **Experimental Results**
- Conclusion

Synthetic Experiments on Trifeature Datasets

- We conduct controlled experiments on a synthetic dataset derived from Trifeature to assess the model's capacity to learn uniqueness, redundancy and synergy.

Model	*redundancy*↑	*uniqueness*↑	*synergy*↑
Cross♣	**100.0**	11.6	50.0
Cross+Self♣	99.7	86.9	50.0
FactorCL♣	99.8	62.5	46.5
MAE	$99.8_{\pm 0.11}$	$82.4_{\pm 3.09}$	$50.1_{\pm 0.24}$
CoMM	$99.9_{\pm 0.06}$	$86.8_{\pm 2.99}$	$71.4_{\pm 3.47}$
InfMasking (ours)	$99.9_{\pm 0.09}$	$\mathbf{90.6}_{\pm 2.31}$	$\mathbf{77.0}_{\pm 4.22}$

♣ denotes results are from "What to align in multimodal contrastive learning?"

Experiments with 2 Modalities on Multibench

- We further evaluate the performance of our model on several real-world multimodal datasets provided by Multibench.

Model	*Regression*	*Classification*				
	V&T EE↓	*MIMIC*↑	*MOSI*↑	*UR-FUNNY*↑	*MUSTARD*↑	**Average*** ↑
Cross♣	$33.09_{\pm 3.67}$	$66.7_{\pm 0.1}$	$47.8_{\pm 1.8}$	$50.1_{\pm 1.9}$	$53.5_{\pm 2.9}$	54.52
Cross+Self♣	$7.56_{\pm 0.31}$	$65.49_{\pm 0.0}$	$49.0_{\pm 1.1}$	$59.9_{\pm 0.9}$	$53.9_{\pm 4.0}$	57.07
FactorCL♣	$10.82_{\pm 0.56}$	$67.3_{\pm 0.0}$	$51.2_{\pm 1.6}$	$60.5_{\pm 0.8}$	$55.80_{\pm 0.9}$	58.7
CoMM	$7.96_{\pm 2.13}$	$66.4_{\pm 0.41}$	$63.7_{\pm 2.5}$	$63.3_{\pm 0.51}$	$64.4_{\pm 1.1}$	64.45
InfMasking (ours)	$\mathbf{4.23}_{\pm 0.37}$	$\mathbf{68.1}_{\pm 0.42}$	$\mathbf{69.0}_{\pm 1.2}$	$\mathbf{64.3}_{\pm 0.9}$	$\mathbf{66.8}_{\pm 2.5}$	**67.05**

♣ denotes results are from “What to align in multimodal contrastive learning?”

Experiments with 3 Modalities on Multibench

- Besides the 2 modalities experiments, we further conducted experiments on the 3 modalities dataset.

Model	*#Mod.*	*V&T CP*↑	*UR-FUNNY*↑
Cross	2	$86.3_{\pm 0.25}$	50.1♣
Cross+Self	2	$87.6_{\pm 0.26}$	59.9♣
CoMM	2	$85.3_{\pm 0.84}$	$63.3_{\pm 0.51}$
InfMasking (ours)	2	$88.5_{\pm 0.33}$	$64.3_{\pm 0.9}$
CMC♣	3	94.1	59.2
CoMM	3	$\mathbf{94.1}_{\pm 0.17}$	$64.8_{\pm 1.13}$
InfMasking (ours)	3	$94.1_{\pm 0.09}$	$\mathbf{65.6}_{\pm 1.15}$

♣ denotes results are from “What to align in multimodal contrastive learning?”

Experiments with 2 Modalities on Multimodal IMDb

- Multimodal IMDb is a real-world multimodal, multi-label dataset designed for movie genre classification. It poses two major challenges: significant class imbalance with genres such as comedy and drama dominating the label distribution, and substantial semantic discrepancy between visual (poster) and textual (plot's description) modalities.

Model	*Modalities*	*weighted-f1*↑	*macro-f1*↑
SimCLR♣	V	$40.35_{\pm 0.23}$	$27.99_{\pm 0.33}$
	V	51.5	40.8
CLIP♣	L	51.0	43.0
	V+L	58.9	50.9
SLIP♣	V+L	$56.54_{\pm 0.19}$	$47.35_{\pm 0.27}$
CLIP♣	V+L	$54.49_{\pm 0.19}$	$44.94_{\pm 0.30}$
CoMM(CLIP backbone)	V+L	$61.29_{\pm 0.73}$	$53.79_{\pm 0.22}$
InfMasking(ours, CLIP backbone)	V+L	$\mathbf{62.60}_{\pm 0.26}$	$\mathbf{55.93}_{\pm 0.19}$

♣ denotes results are from "What to align in multimodal contrastive learning?"

Outline

- Background
- Method
- Experimental Results
- **Conclusion**

Conclusion

We introduce a contrastive synergistic information extraction method via infinite masking.

- InfMasking stochastically occludes a substantial proportion of features from each modality during the fusion process. This masking preserves only partial information, creating fused representations with varied synergistic patterns
- Unmasked fused representations are aligned with these masked ones via mutual information maximization to encode comprehensive synergistic information.
- To address the expensive computation of mutual information estimates with infinite masking, we derive an InfMasking loss to approximate the calculation of this loss function.

THANKS