What to Align in Multimodal Contrastive Learning?

B. Dufumier[1,2,*], J. Castillo Navarro[1,3,*], D. Tuia[1], J-P. Thiran[1]

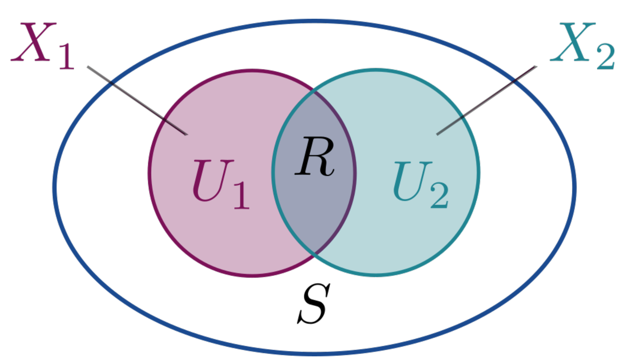
[1]École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
[2] NeuroSpin, CEA Saclay - Université Paris-Saclay, France
[3] Conservatoire National des Arts et Métiers, CEDRIC laboratory, France
[*] denotes equal contributions

EPFL · NeuroSpin · le cnam · ICLR International Conference On Learning Representations

1. Motivation

- **Humans experience the world through multi-sensory integration**, blending information across multiple modalities.

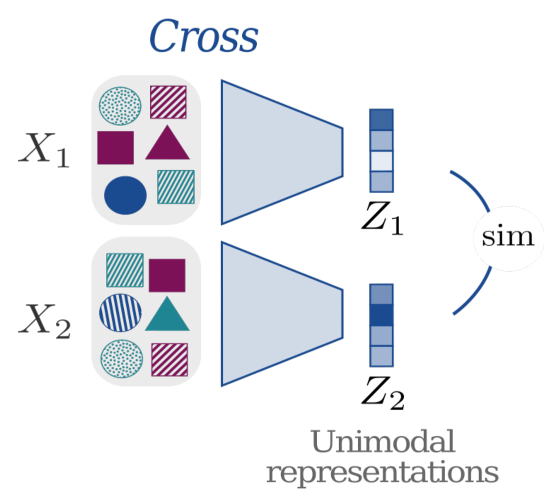
- **Multimodal representation learning** preserves:
  → modality-specific information (**U**niqueness)
  → shared semantics (**R**edundancy)
  → cross-modal synergy (**S**ynergy)
- How to model these quantities ?
  ⇝ **Partial information decomposition** (PID)



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

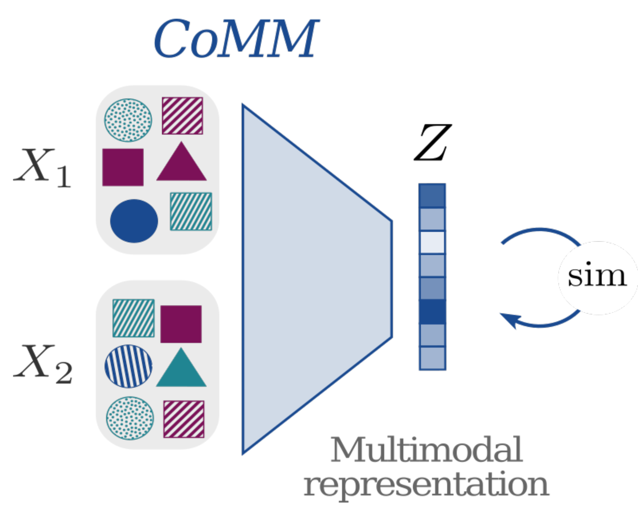
Can we capture multimodal interactions in a self-supervised way?

2. Beyond cross-modal alignment

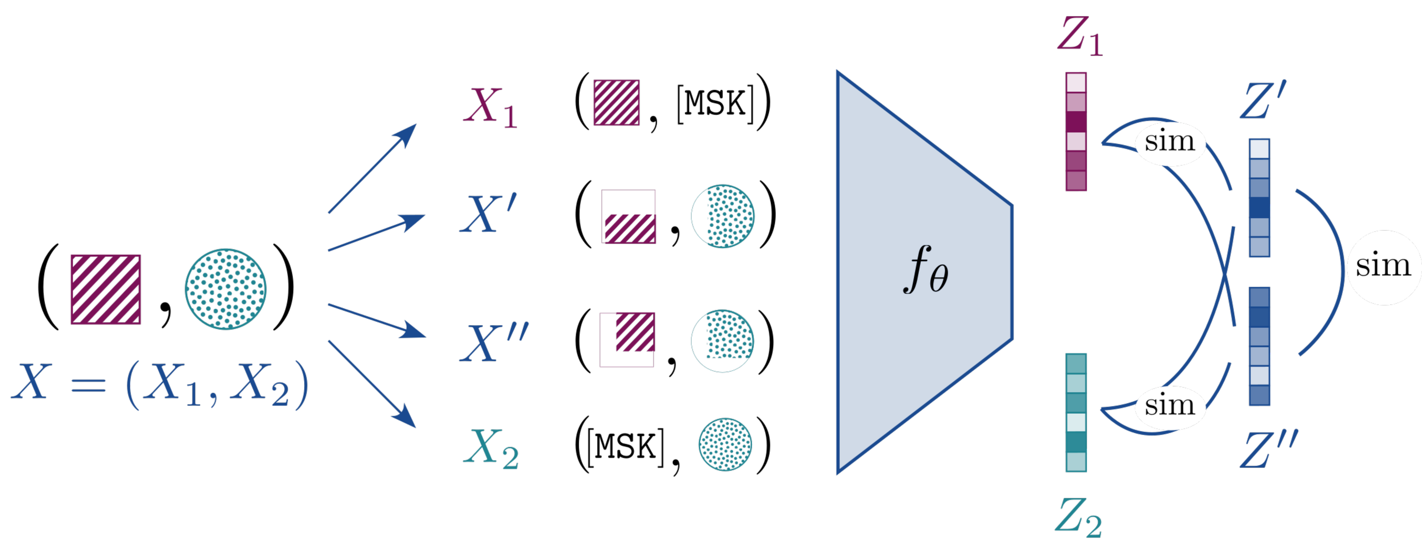


- CLIP-like models align representations from two modalities
- It only learns **redundant** information, neglecting other interactions

- CoMM encodes multiple modalities to a **single multimodal space**
- It aligns **multimodal representations**, integrating redundant, unique and synergistic interactions.



3. CoMM



CoMM's training

Given a set of *minimal label preserving multimodal augmentations* $\mathcal{T}^\star$

- Draw $t', t'' \in \mathcal{T}^\star$ to obtain X' and X''
- Get projections X_1 and X_2
- Get multimodal embeddings Z', Z'' and Z_1, Z_2
- Contrastive loss: $\mathcal{L}_{\text{CoMM}}$

Loss function

- $\mathcal{L} = -\hat{I}_{\text{NCE}}(Z', Z'')$
- $\mathcal{L}_i = -\frac{1}{2}\left(\hat{I}_{\text{NCE}}(Z_i, Z') + \hat{I}_{\text{NCE}}(Z_i, Z'')\right)$

⇝ $\boxed{\mathcal{L}_{\text{CoMM}} = \mathcal{L} + \sum_{i=1}^{n} \mathcal{L}_i}$

Theoretical guarantees

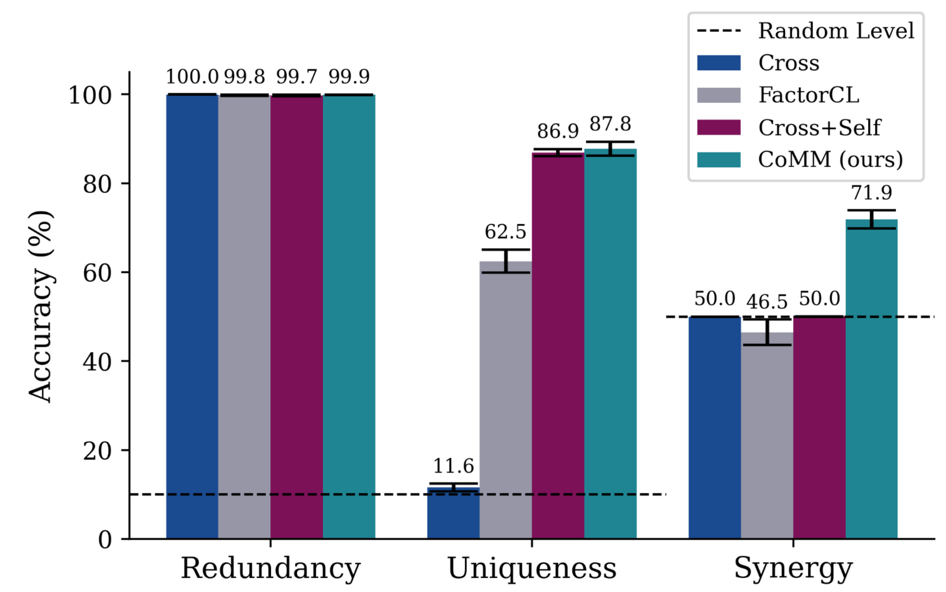
Lemma 2. By optimizing f_θ to maximize $I(Z_\theta; Z'_\theta)$, and if we assume an expressive enough network f_θ, we have at optimum: $I(Z_{\theta^\star}, Z'_{\theta^\star}) = I(X, X')$

Lemma 3. Let $f_{\theta^\star}$ be optimal, i.e. $f_{\theta^\star}$ maximizes $I(Z_\theta, Z'_\theta)$. Then, we have the equality $I(Z'_{\theta^\star}; Y) = I(X'; Y)$. If we consider the special case $\mathcal{T} = \{t_i\}$ such that $X' = t_i(X) = X_i$ and $Z'_{\theta^\star} = f_{\theta^\star}(X_i) = Z_i$ for $i \in \{1, 2\}$, then it follows: $I(Z_i; Y) = I(X_i; Y) = R + U_i$

4. Controlled experiments: Bimodal trifeatures

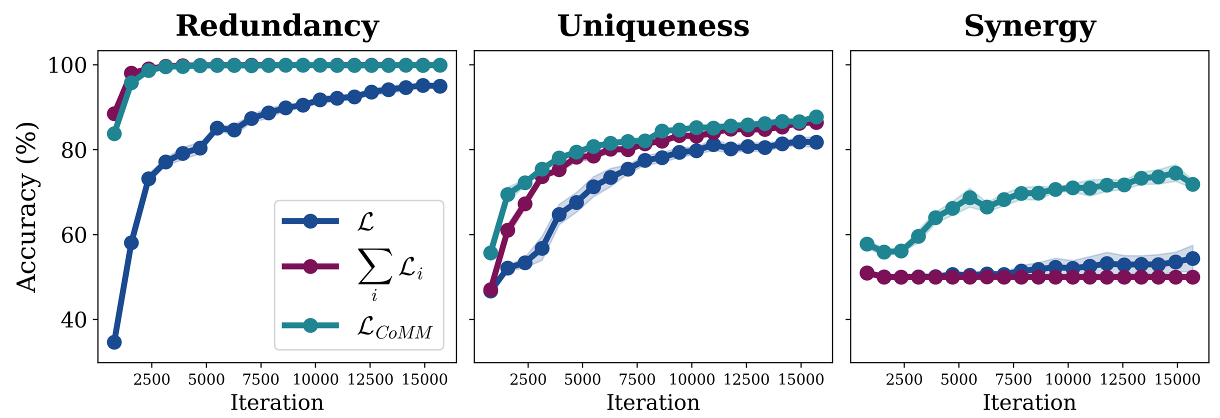
→ 2 streams of trifeature samples
→ 3 features: color, shape and texture, 10 of each

- **Uniqueness.** Given a pair with **different textures**:
  → U_i: predict the i-th texture
- **Redundancy.** Given a pair with **same shape**:
  → **R**: predict the shape of inputs
- **Synergy.** Given a **unique matching** (texture, color) & a pair of samples:
  → **S**: matching satisfied?



⇝ **CoMM is the only model to learn synergy!**

Ablation study on the loss function



- $\sum_i \mathcal{L}_i$ learns redundancy and uniqueness, but fails at synergy
- $\mathcal{L}$ learns all the terms, but slowly
- **$\mathcal{L}_{\text{CoMM}}$ is the perfect compromise**

5. Results with 2 modalities

- **MM-IMDb**
  → Modalities: Images & Text (movie poster + description)
  → Task: Multi-label classification (movie genre)

  ⇝ **CoMM beats modern vision-language models!**

Model	*Mod.*	*w-f1*	*m-f1*
CLIP	V	51.5	40.8
	L	51.0	43.0
	V+L	58.9	50.9
BLIP-2	V+L	57.4	49.9
CoMM (CLIP init)	V+L	61.4	54.6
CoMM (BLIP-2 init)	V+L	**64.7**	**58.4**
MFAS	V+L	62.5	55.6
CoMM† (CLIP init)	V+L	64.9	58.9
CoMM† (BLIP-2 init)	V+L	**67.3**	**62.0**
LLaVA-NeXT	V+L	64.2	56.5

Rows in color are supervised. †: supervised fine-tuning.

- **MultiBench**
  → Diverse data modalities: tabular, time-series, text, images, etc.
  → **Complex multimodal scenarios**: varying degrees of shared and unique relevant information.

Model	*Regression*	*Classification*			
	V&T EE ↓	*MIMIC* ↑	*MOSI* ↑	*UR-FUNNY* ↑	*MUsTARD* ↑
Cross	33.0	66.7	47.8	50.1	53.5
Cross+Self	7.5	65.4	49.0	59.9	53.9
FactorCL	10.8	**67.3**	51.2	60.5	55.8
CoMM	**4.5**	66.4	**67.5**	**63.1**	**63.9**
SupCon	-	67.4	47.2	50.1	52.7
FactorCL-SUP	1.7	**76.8**	69.1	63.5	69.9
CoMM (fine-tuned)	**1.3**	68.1	**74.9**	**65.9**	**70.4**

⇝ **CoMM is a versatile and efficient multimodal model**

6. Results with 3 modalities

- CoMM can be trained with **more than 2 modalities!**

Model	#Mod.	*V&T CP*	*UR-FUNNY*
Cross	2	84.4	50.1
Cross+Self	2	86.8	59.9
CoMM (ours)	2	88.1	63.1
CMC	3	94.1	59.2
CoMM (ours)	3	**94.2**	**64.6**

→ **Consistent improvement** with a third modality.

7. Perspectives

Visit our website!

- PID theory is limited to 2 modalities
  ⇝ Extension using O-Information
- Interpretability of CoMM
  ⇝ Disentangle multimodal interactions
- Data augmentation computational cost
  ⇝ Investigate knowledge distillation