Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 3.2: Multimodal Coordination and Fission

Paul Liang

** Co-lecturer: Louis-Philippe Morency. Original course co-developed with Tadas Baltrusaitis. Spring 2021 and 2022 editions taught by Yonatan Bisk.*

Objectives of today's class

- Representation coordination
  - Coordination functions
    - Kernel similarity functions
    - Canonical correlation analysis
  - Contrastive learning
  - Information, entropy and mutual information
- Representation fission
  - Factorized multimodal representations
  - Clustering and fine-grained fission

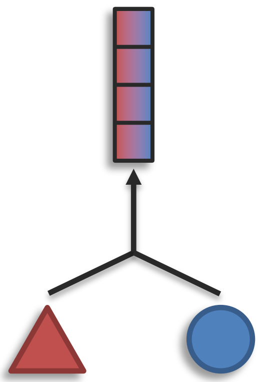
Multimodal Representation

Challenge 1: Representation

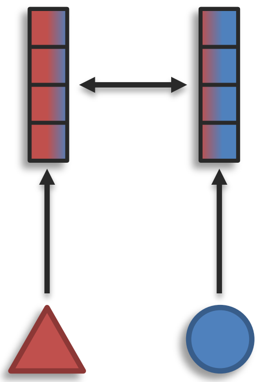
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

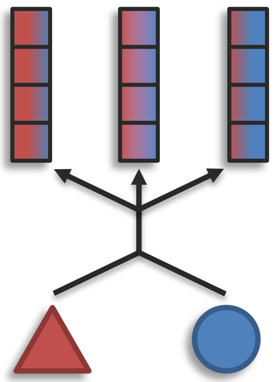
Fusion



Coordination

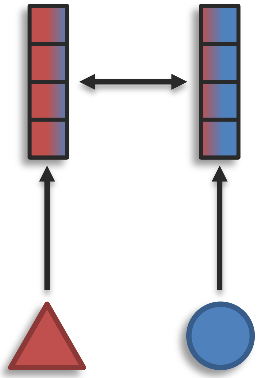


Fission



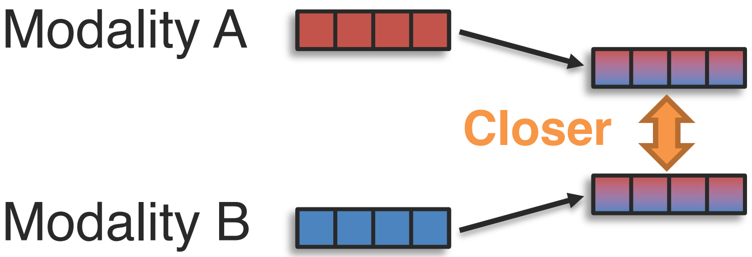
Representation Coordination

Sub-Challenge 1b: Representation Coordination

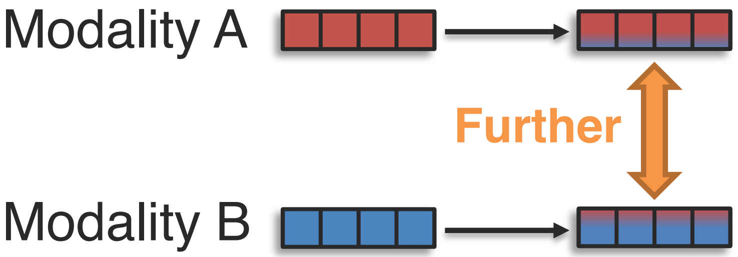


Definition: Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions

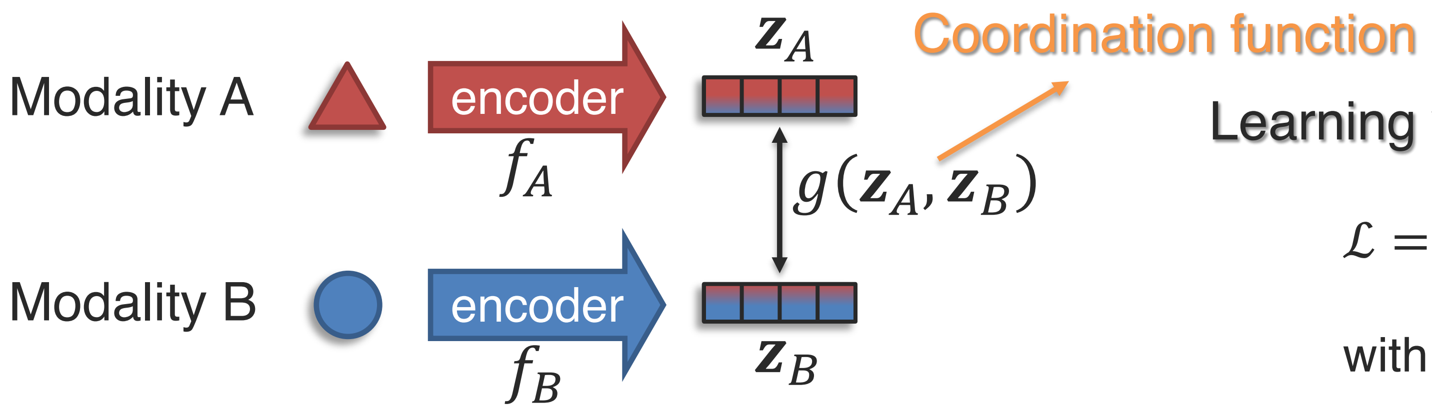
Strong Coordination:



Partial Coordination:



Coordination Function



Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

➡ Requires paired data

Examples of coordination function:

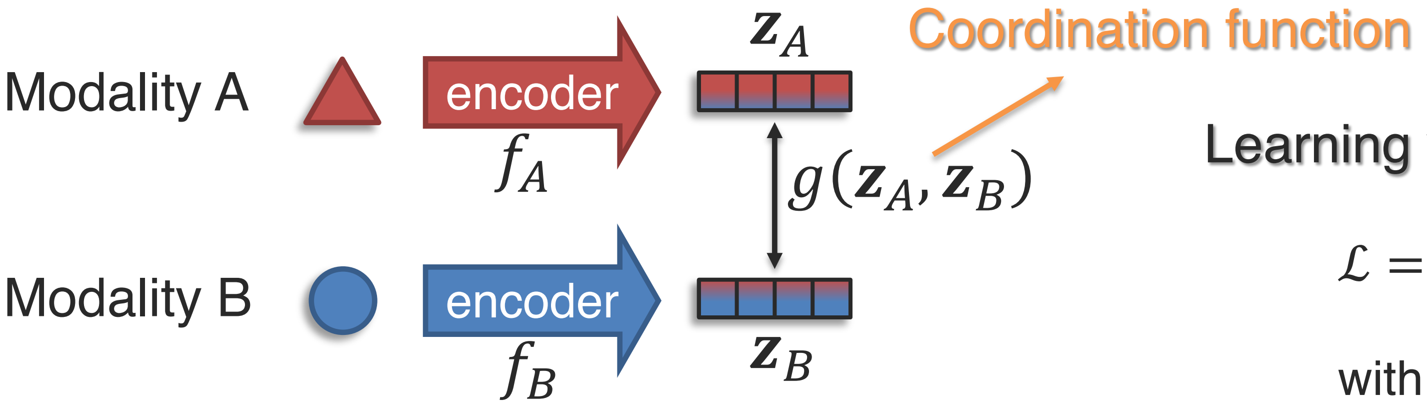
① Cosine similarity:

$$g(\mathbf{z}_A, \mathbf{z}_B) = \frac{\mathbf{z}_A \cdot \mathbf{z}_B}{\|\mathbf{z}_A\| \|\mathbf{z}_B\|}$$

Strong coordination!

➡ For normalized inputs (e.g., $\mathbf{z}_A - \overline{\mathbf{z}_A}$) , equivalent to *Pearson correlation coefficient*

Coordination Function



Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

② Kernel similarity functions:

$$g(\mathbf{z}_A, \mathbf{z}_B) = k(\mathbf{z}_A, \mathbf{z}_B)$$

- Linear
- Polynomial
- Exponential
- RBF

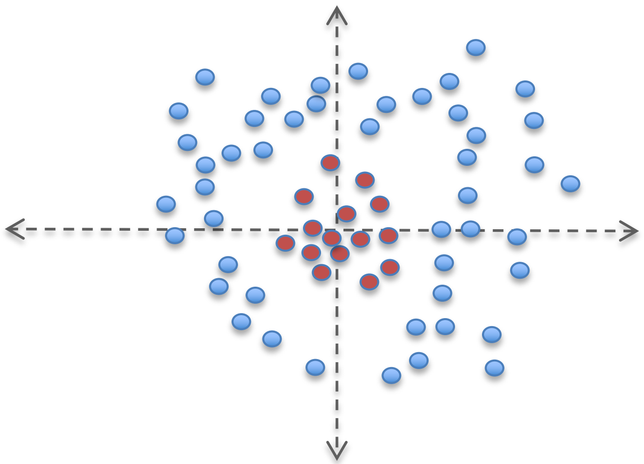
➡ All these examples bring relatively strong coordination between modalities

Kernel Function

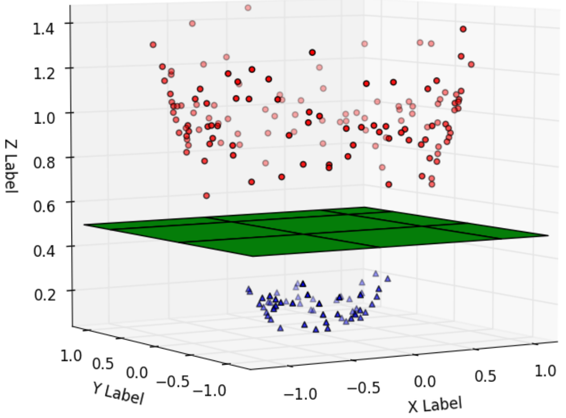
A kernel function: Acts as a similarity metric between data points

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j) \rangle$$

➡ $\phi(\boldsymbol{x})$ can be high-dimensional space!



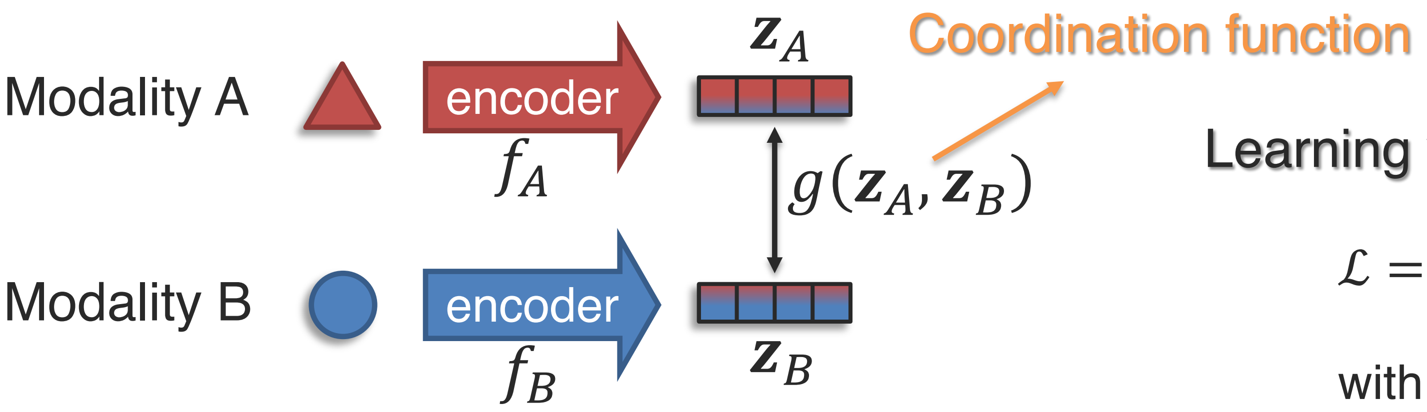
Not linearly separable in x space



Same data, but now linearly separable in $\phi(\boldsymbol{x})$ space

➡ Radial Basis Function (RBF) Kernel : $K(x_i, x_j) = \exp - \frac{1}{2\sigma^2} \left\| x_i - x_j \right\|^2$

Coordination Function



Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\triangle), f_B(\bigcirc)\big)$$

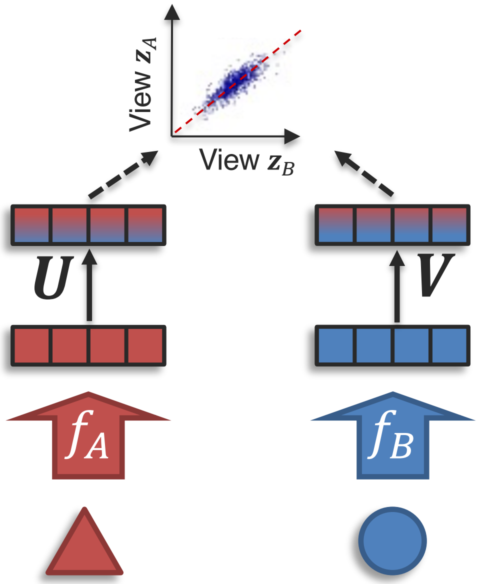
with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

③ Canonical Correlation Analysis (CCA):

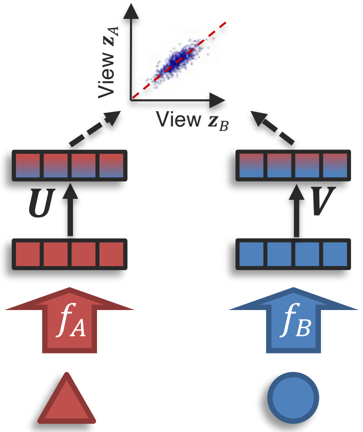
$$\underset{V,U,f_A,f_B}{\operatorname{argmax}}\; corr(\mathbf{z}_A, \mathbf{z}_B)$$

➡ CCA includes multiple projections, all orthogonal with each others

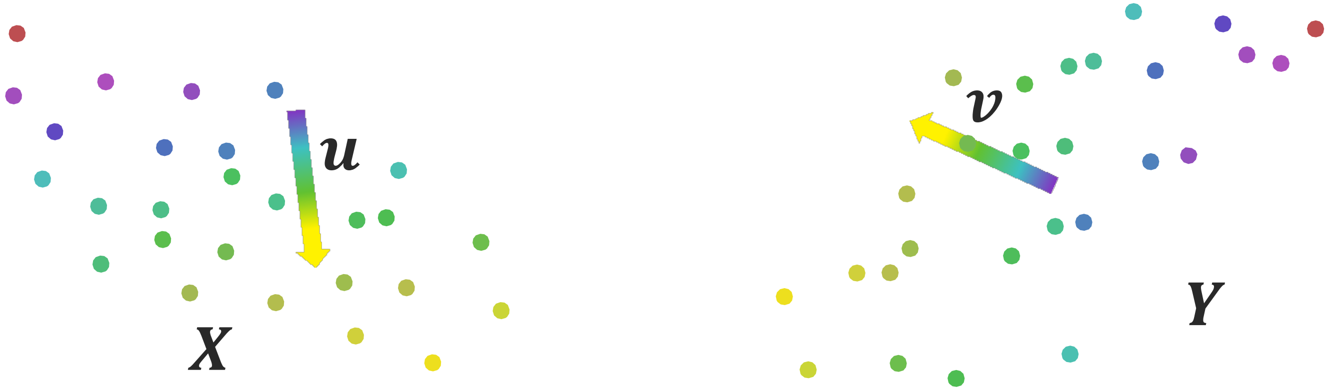


Correlated Projection

1. Learn two linear projections, one for each view, that are maximally correlated:



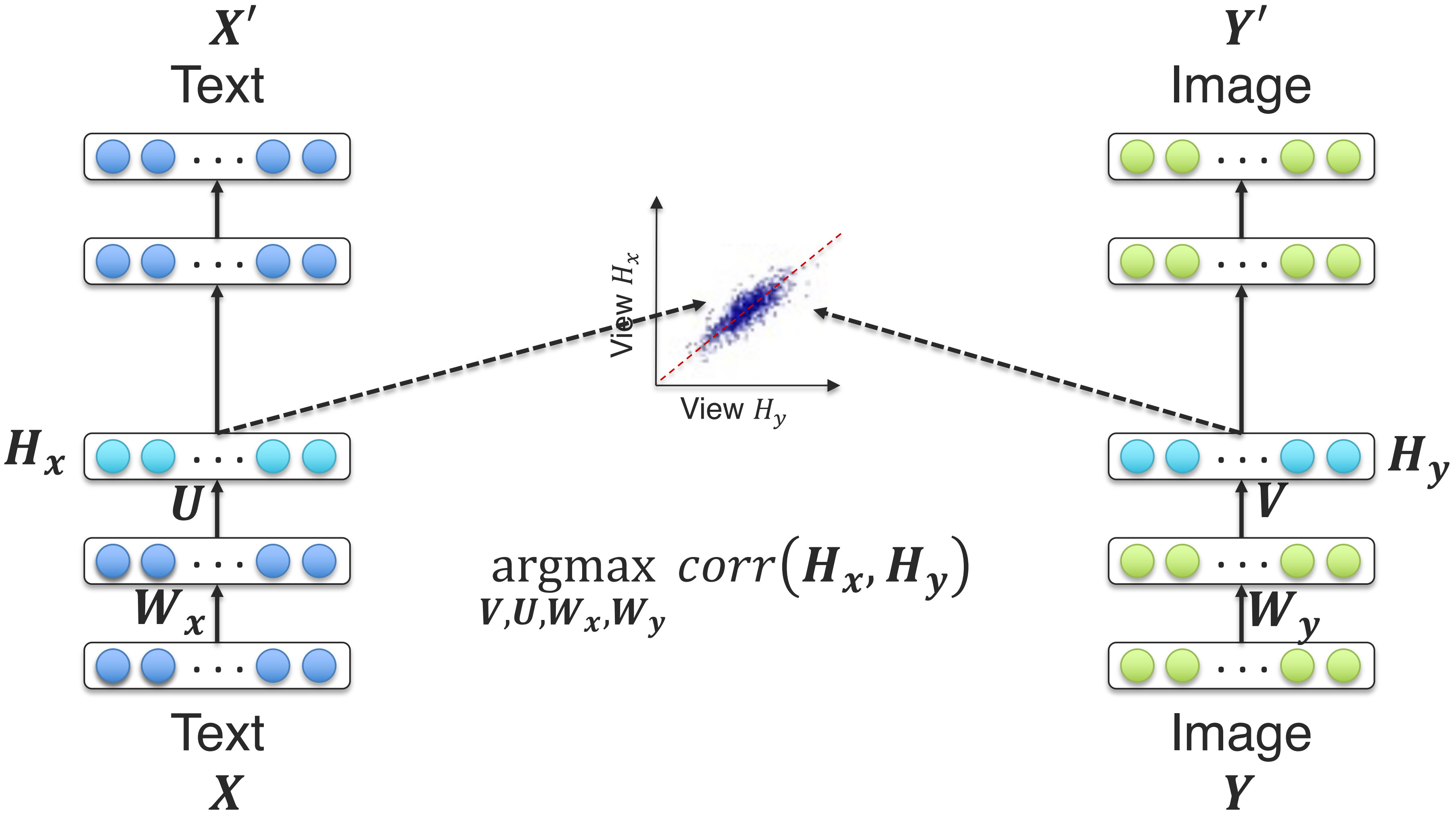
$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr\left(\boldsymbol{u}^T \boldsymbol{X}, \boldsymbol{v}^T \boldsymbol{Y}\right)$$



Two views $\boldsymbol{X}, \boldsymbol{Y}$ where same instances have the same color

Remember that X and Y consist of paired data

Deep Canonically Correlated Autoencoders (DCCAE)

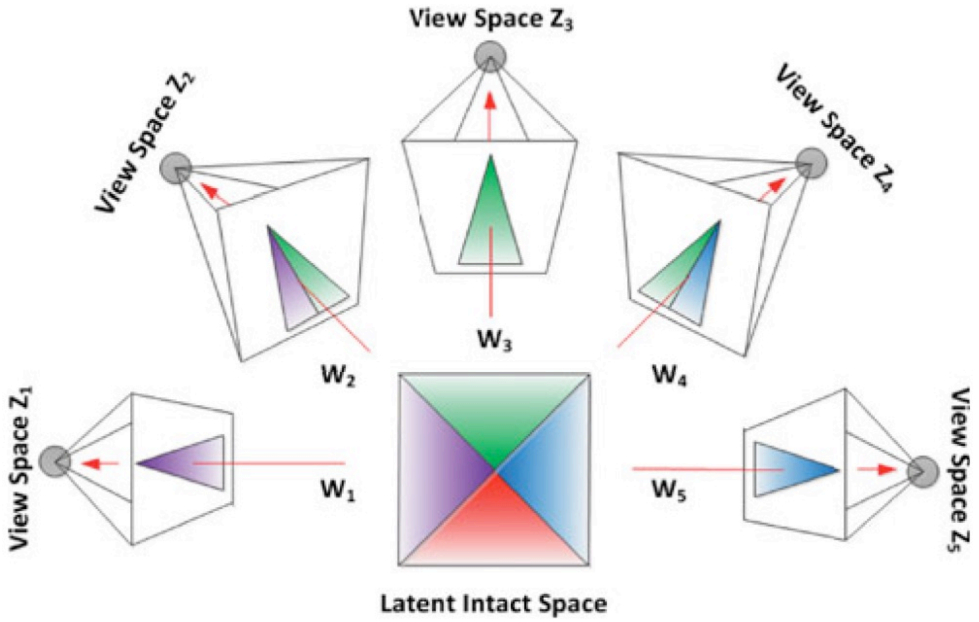


$$\underset{V,U,W_x,W_y}{\text{argmax}} \; corr(H_x, H_y)$$

Wang et al., On deep multi-view representation learning, PMLR 2015

Multi-view Latent “Intact” Space

Given multiple views z_i from the same “object”:

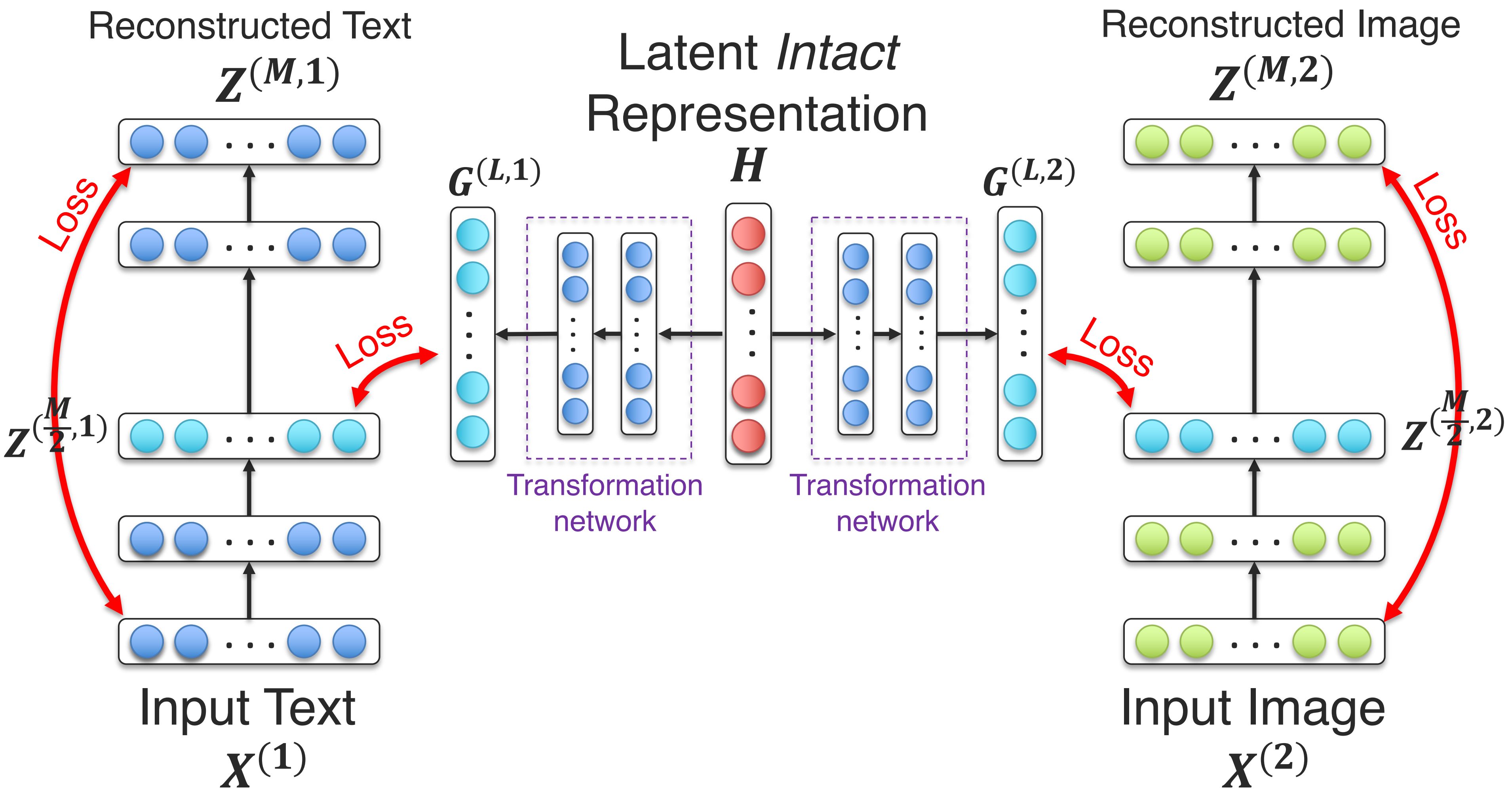


1) There is an “intact” representation which is *complete* and *not damaged*

2) The views z_i are partial (and possibly degenerated) representations of the intact representation

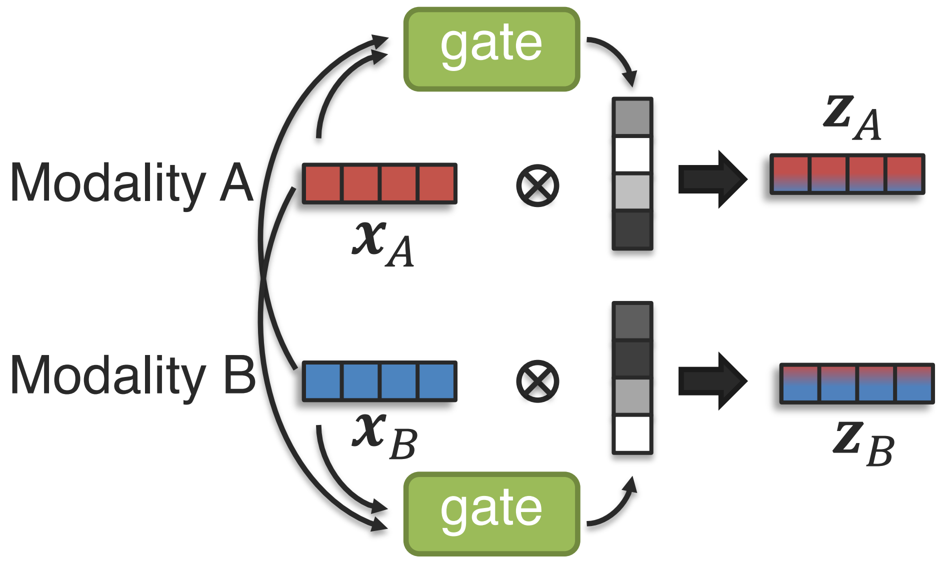
Xu et al., Multi-View Intact Space Learning, TPAMI 2015

Auto-Encoder in Auto-Encoder Network



Zhang et al., AE2-Nets: Autoencoder in Autoencoder Networks, CVPR 2019

Gated Coordination



Gated coordination:

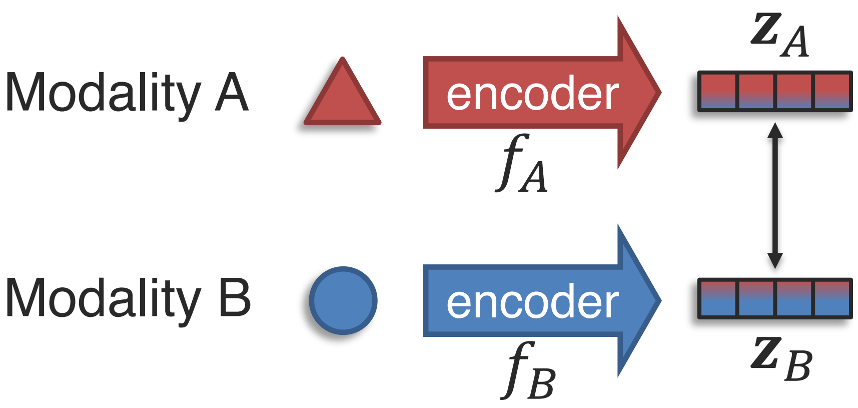
$$z_A = g_A(x_A, x_B) \cdot x_A$$

$$z_B = g_B(x_A, x_B) \cdot x_B$$

→ Related to attention modules in transformers

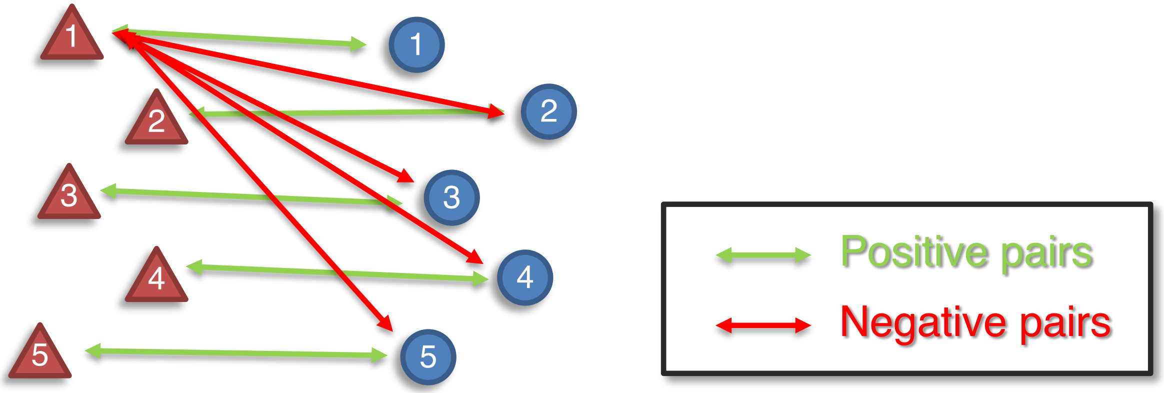
More about it next week!

Coordination with Contrastive Learning



Paired data: { △ , ● }

(e.g., images and text descriptions)



Contrastive loss:

➡ brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

$$\max\{0, \alpha + sim(\boldsymbol{z}_A, \boldsymbol{z}_B^{+}) - sim(\boldsymbol{z}_A, \boldsymbol{z}_B^{-})\}$$

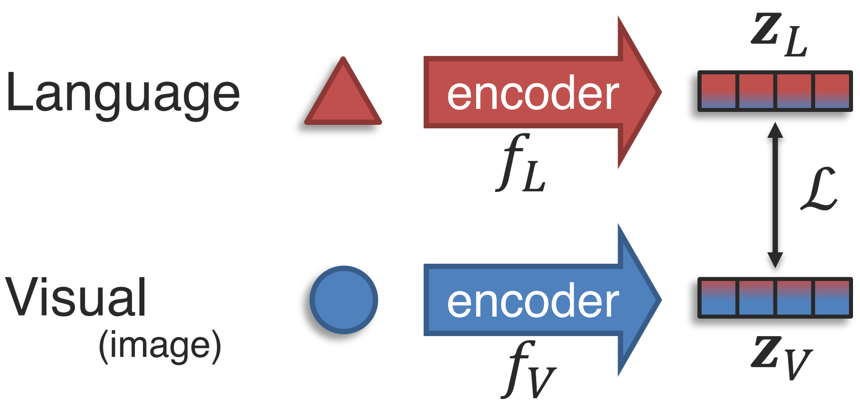
positive pairs

negative pair

Similarity functions are often cosine similarity

➡ Similar to hinge loss

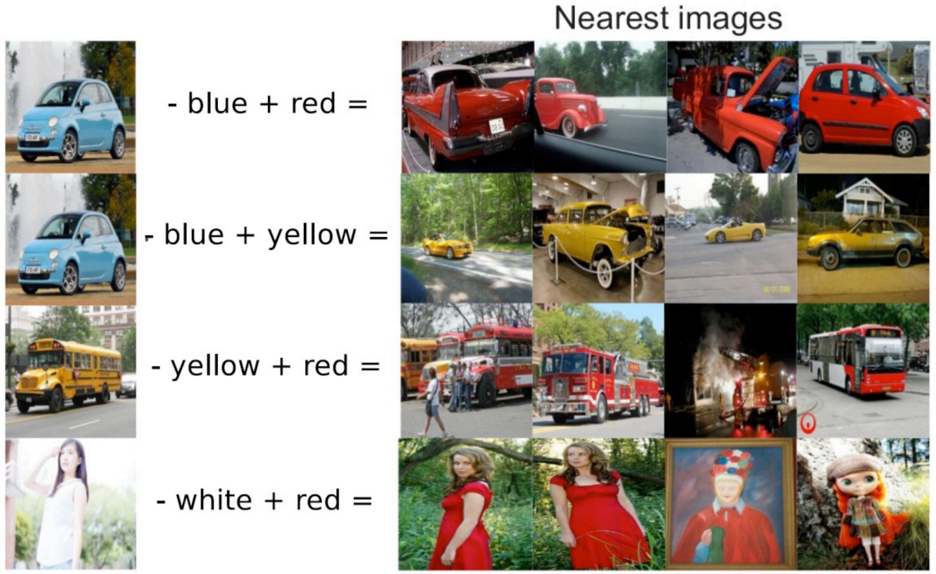
Example – Visual-Semantic Embeddings

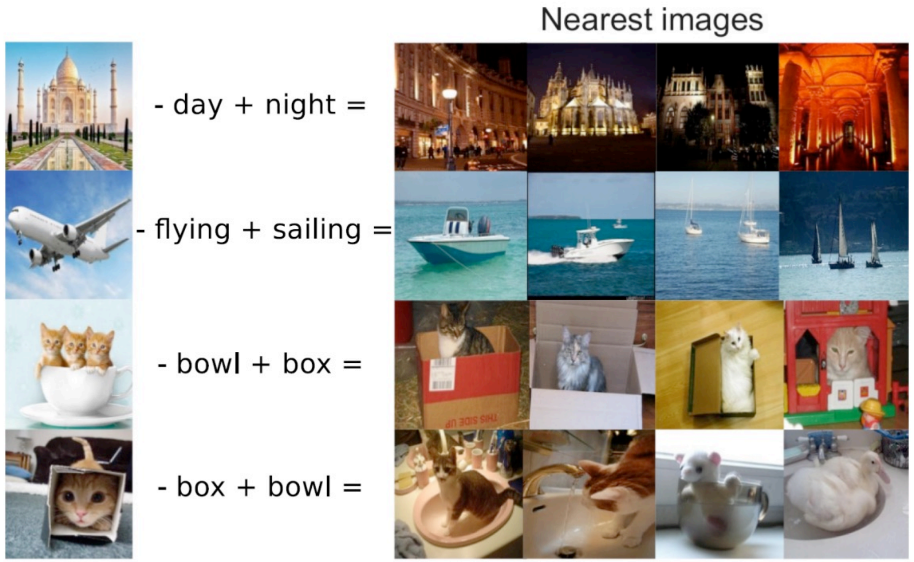


Two contrastive loss terms:

$$\max\{0, \alpha + sim(\boldsymbol{z}_L, \boldsymbol{z}_V^+) - sim(\boldsymbol{z}_L, \boldsymbol{z}_V^-)\}$$

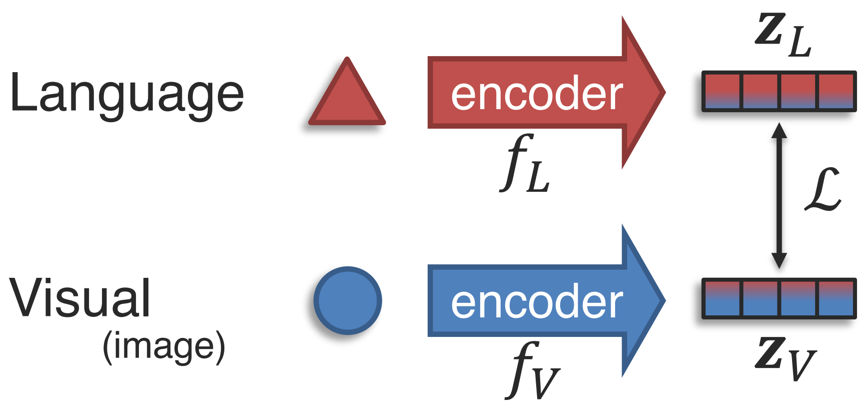
$$+ \max\{0, \alpha + sim(\boldsymbol{z}_V, \boldsymbol{z}_L^+) - sim(\boldsymbol{z}_V, \boldsymbol{z}_L^-)\}$$





Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, NIPS 2014

Example – CLIP (Contrastive Language–Image Pre-training)



Popular contrastive loss: InfoNCE

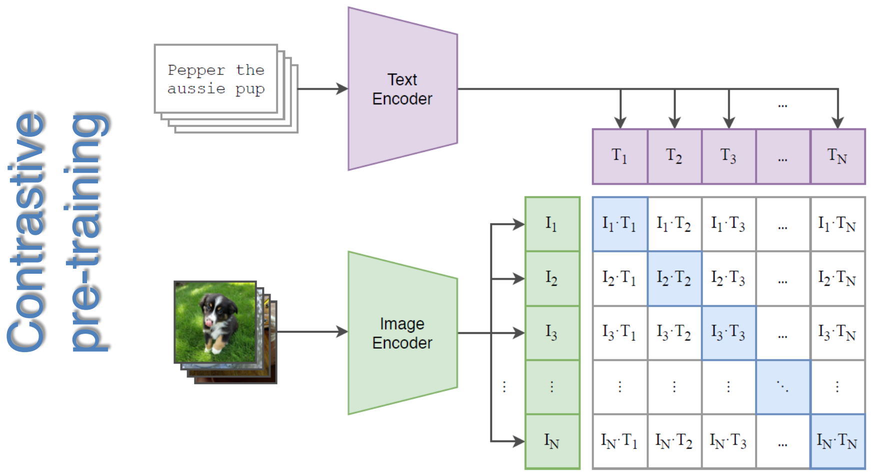
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N}\text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

positive pairs

Similarity function can be cosine similarity

negative pairs and positive pairs

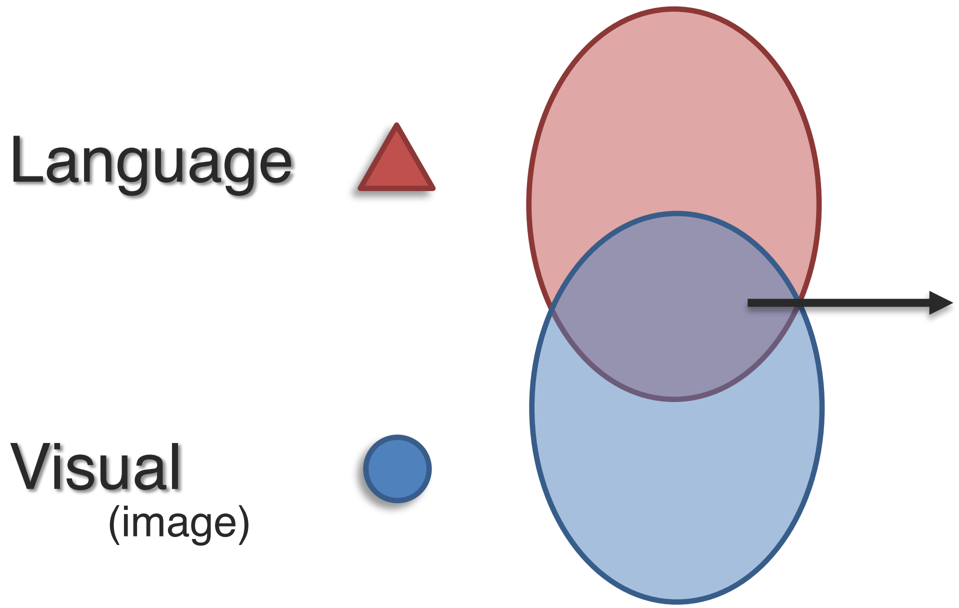
Positive and negative pairs:



- CLIP encoders (f_L and f_V) are great for language-vision tasks
- $\boldsymbol{z}_L$ and $\boldsymbol{z}_V$ are coordinated but not identical representation spaces

[Radford et al., Learning Transferable Visual Models From Natural Language Supervision. ICML 2021]

Multimodal Coordination – Information Theory

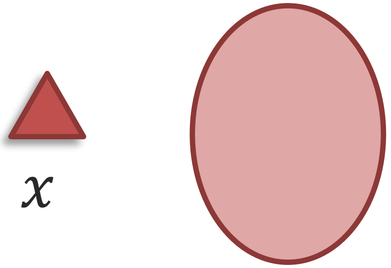


Information in both modalities

- Described people, objects, actions
- Illustrative gestures, motion
- …

Information and Entropy – Information Theory

Language x



How much information in the modality?

Information Theory **(Shannon, 1948)**

Main intuition: "Information value" of a communicated message x depends on how random its content is

x: "1,1,1,1,1,1,1,1,1,1,1,1"

➡ Not very random… So, low information

x: "0,1,0,1,0,0,1,1,1,0,0,1"

➡ More random… So, higher information

Information content $I(x)$

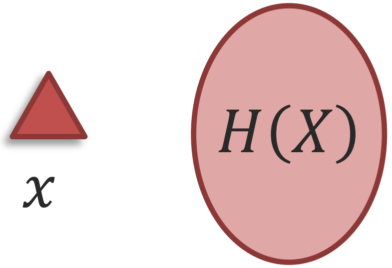
$$I(x) \sim \frac{1}{p(x)}$$

$$I(x) = \log\left(\frac{1}{p(x)}\right) = -\log(p(x))$$

Shannon, A Mathematical Theory of Communication, 1948

Information and Entropy – Information Theory

Language



How much information in the modality?

Information Theory (Shannon, 1948)

Information content $I(X) = -\log(p(X))$

- For discrete alphabet $\mathcal{X}$, then X is discrete random variable

Entropy: weighted average of all possible outcomes from $\mathcal{X}$

$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log(p(X))] = -\sum_{x \in \mathcal{X}} p(X)\log(p(X))$$

- Entropy can also be defined for continuous random variables

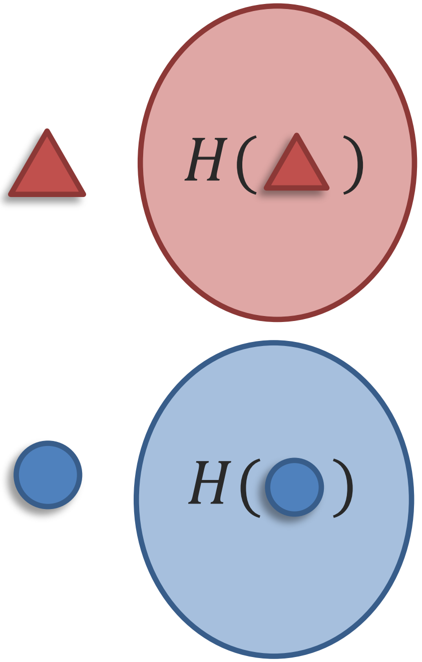
Entropy with Two Modalities

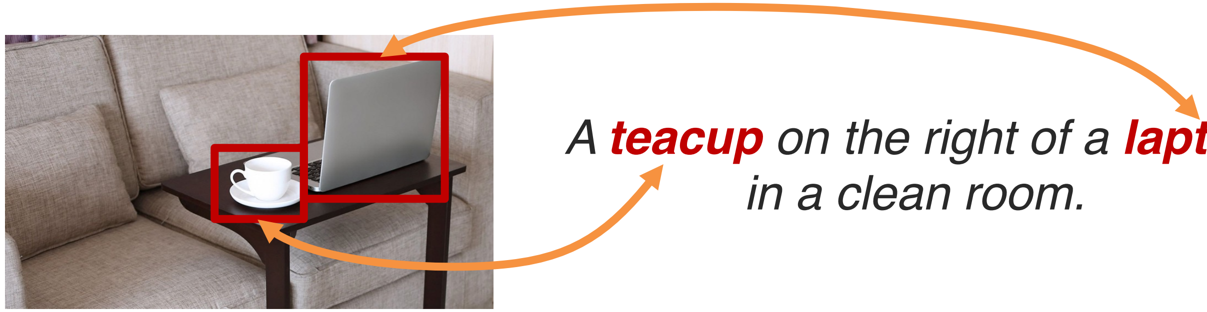
If no overlapping information

But in most real-world scenarios, modalities are *inter-connected*

Modality A

Modality B

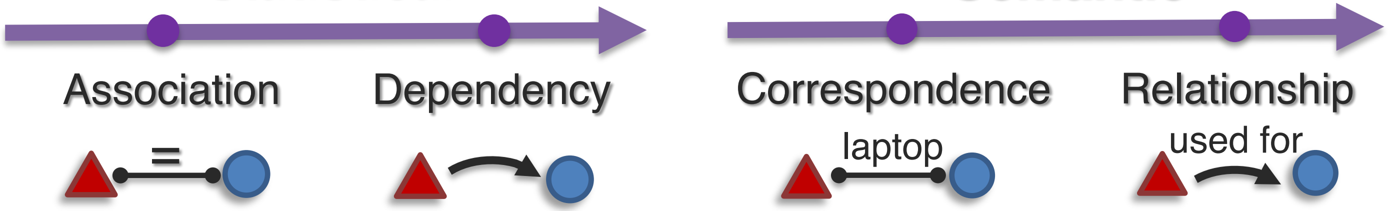




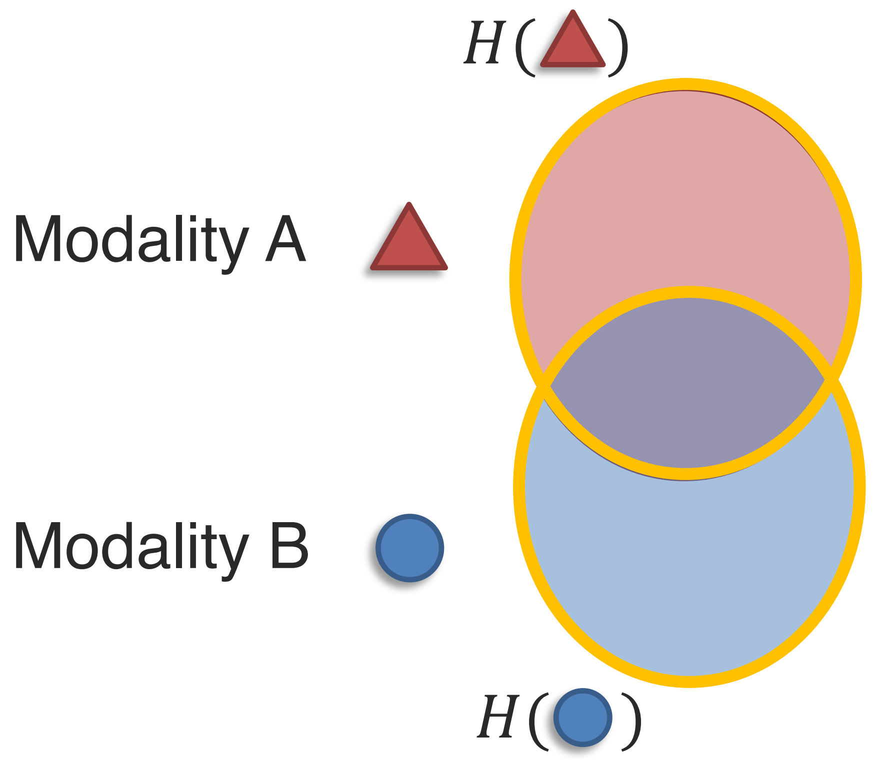
*A **teacup** on the right of a **laptop** in a clean room.*

Statistical

Semantic



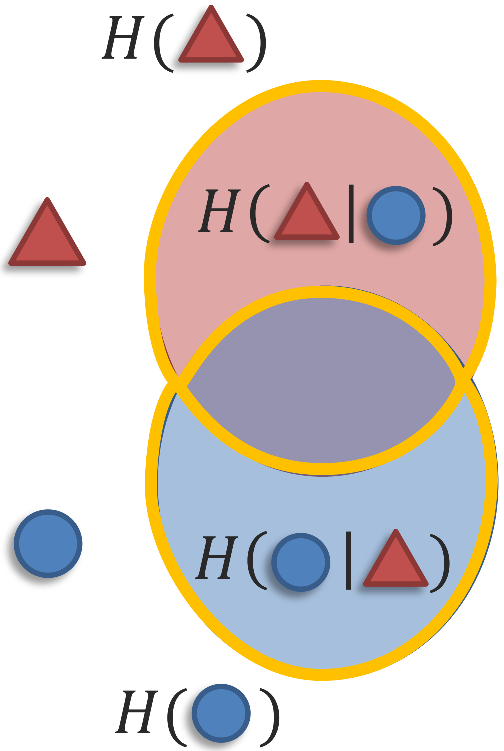
Entropy with Two Modalities



Entropy with Two Modalities

Modality A

Modality B



Conditional entropy $H(Y|X)$

$$H(Y|X) = -\mathbb{E}_{X,Y}[\log p(y|x)]$$

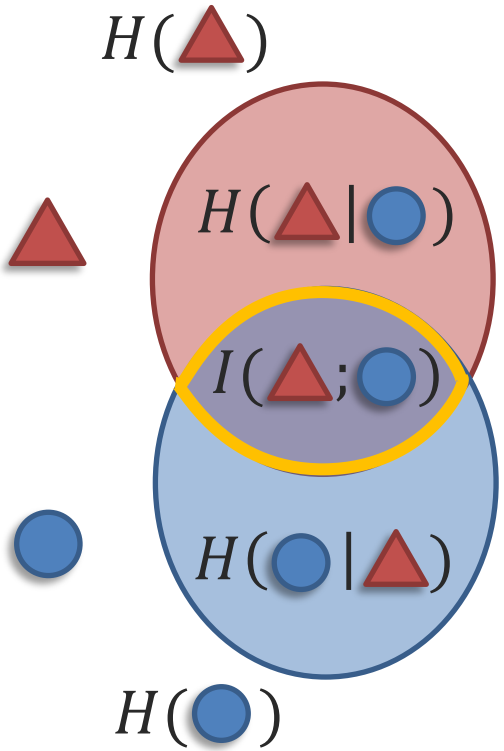
$$= -\mathbb{E}_{X,Y}\left[\log \frac{p(x,y)}{p(x)}\right]$$

If X and Y independent, $H(Y|X) = H(Y)$.
If X fully determines Y, then $H(Y|X) = 0$.

Entropy with Two Modalities

Modality A

Modality B



Mutual information $I(X;Y)$

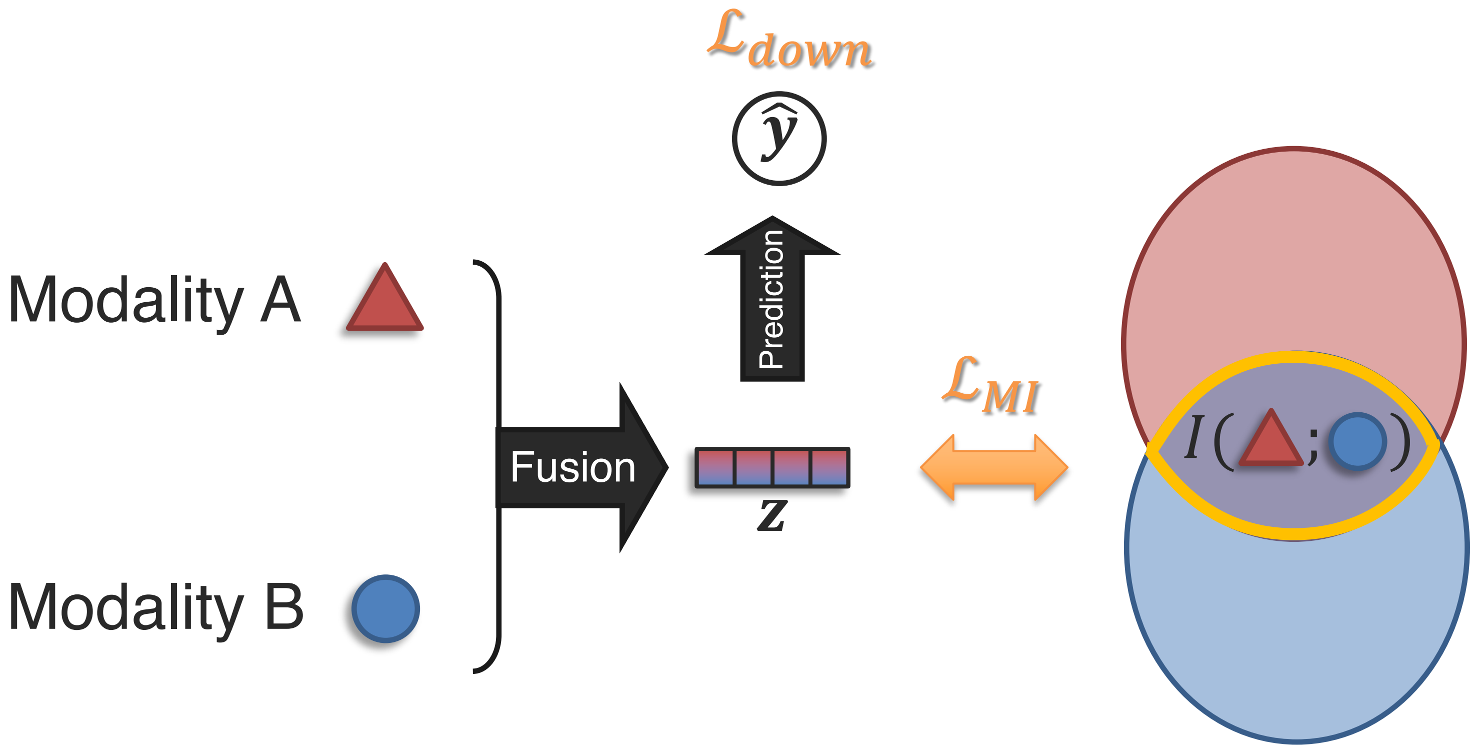
$$I(X;Y) = H(X) - H(X|Y)$$

$$= \mathbb{E}_{X,Y}\left[\log \frac{1}{P_X(x)} + \log \frac{P_{XY}(x,y)}{P_Y(y)}\right]$$

$$I(X;Y) = \mathbb{E}_{X,Y}\left[\log \frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right]$$

using KL-divergence ⇦ $I(X;Y) = D_{KL}(P_{XY}(x,y) \parallel P_X(x)P_Y(y))$

Multimodal Fusion with Mutual Information

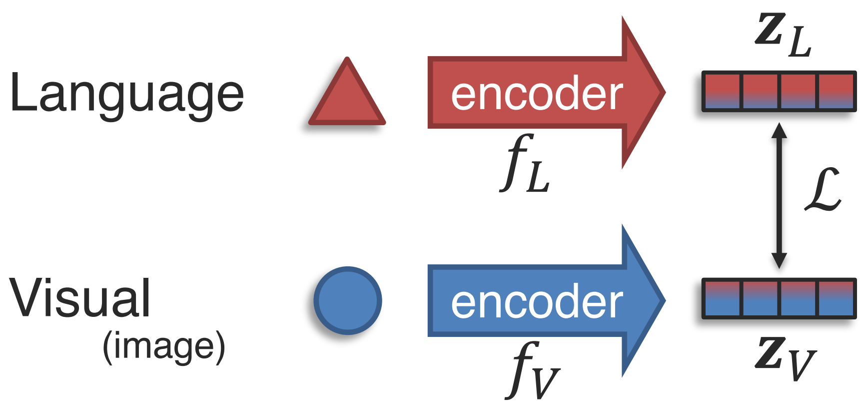


Assumption?

Information present in both modalities is most important for the downstream task

Colombo et al., Improving Multimodal Fusion via Mutual Dependency Maximization, EMNLP 2021

Contrastive Learning and Connected Modalities

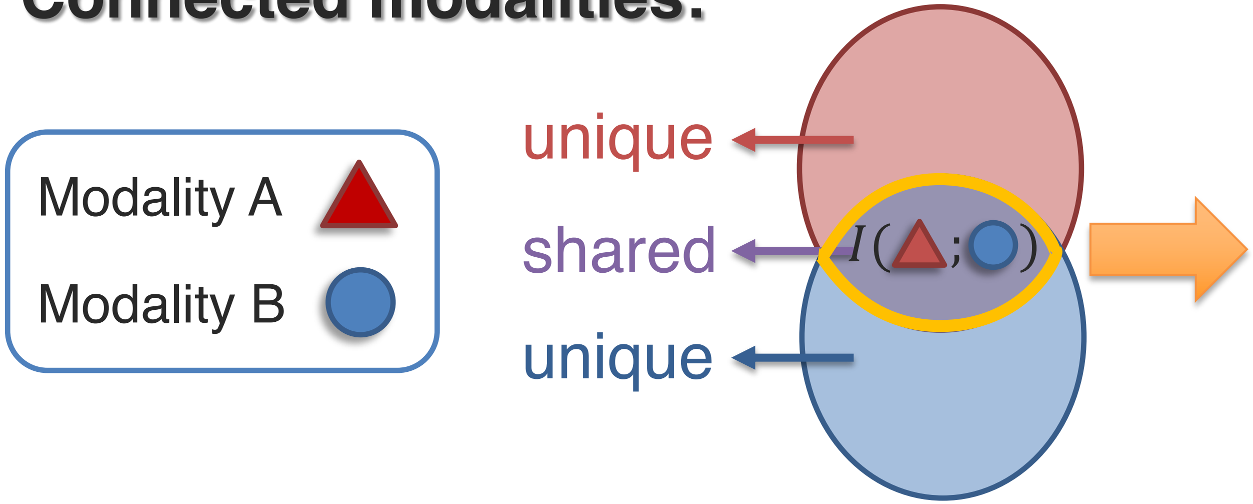


Popular contrastive loss: InfoNCE

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N}\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

CLIP focuses on shared connections

Connected modalities:

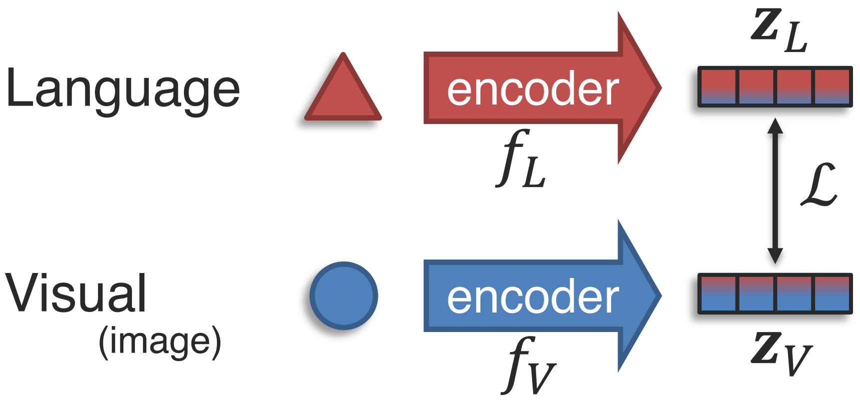


Mutual information $I(X;Y)$

$$\mathbb{E}_{X,Y}\left[\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right]$$

[Oord et al., Representation Learning with Contrastive Predictive Coding. 2018]

Contrastive Learning and Mutual Information



InfoNCE:

$$\mathcal{L} = -\mathbb{E}\left[\log \frac{f(\boldsymbol{x}_A^i, \boldsymbol{x}_B^i)}{\sum_{j=1}^{N} f(\boldsymbol{x}_A^i, \boldsymbol{x}_B^j)}\right]$$

critic function

Critic function f is trained to be a binary classifier distinguishing $\boldsymbol{x}_A, \boldsymbol{x}_B \sim p(\boldsymbol{x}_A, \boldsymbol{x}_B)$ vs $\boldsymbol{x}_A, \boldsymbol{x}_B \sim p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)$

At optimal loss, $f^*(\boldsymbol{x}_A, \boldsymbol{x}_B) = \frac{p(\boldsymbol{x}_A, \boldsymbol{x}_B)}{p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)}$.

Plugging f^* back into $\mathcal{L}$ gives:

$$\mathcal{L}^* \geq \mathbb{E}\left[\log \frac{p(\boldsymbol{x}_A)p(\boldsymbol{x}_B)}{p(\boldsymbol{x}_A, \boldsymbol{x}_B)} N\right] = -I(X_A, X_B) + \log N$$

InfoNCE/CL:

- 'Captures' mutual information
- Optimizes a lower bound on mutual information

In other words: $I(X_A, X_B) \geq \log N - \mathcal{L}^*$

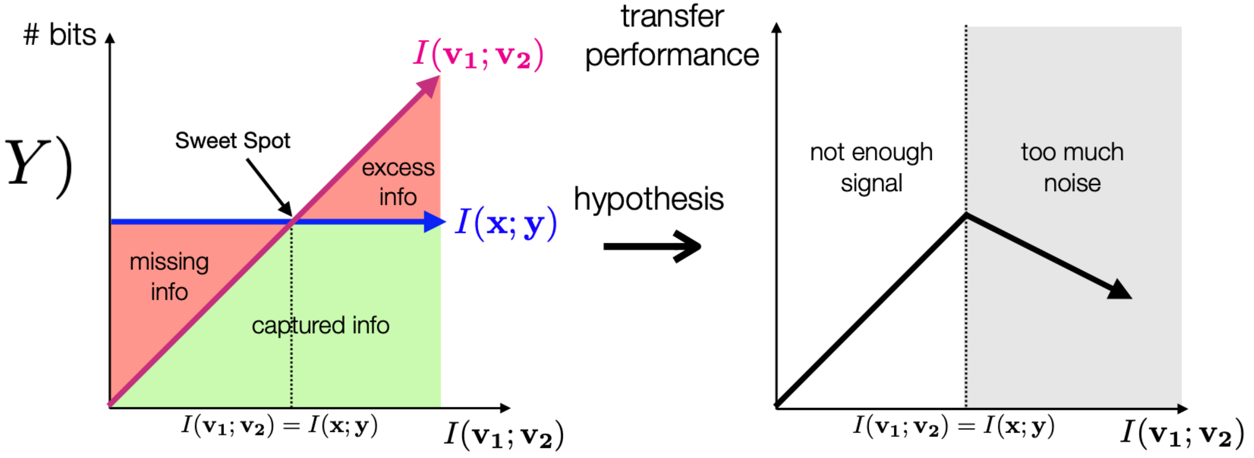
[Oord et al., Representation Learning with Contrastive Predictive Coding. 2018]

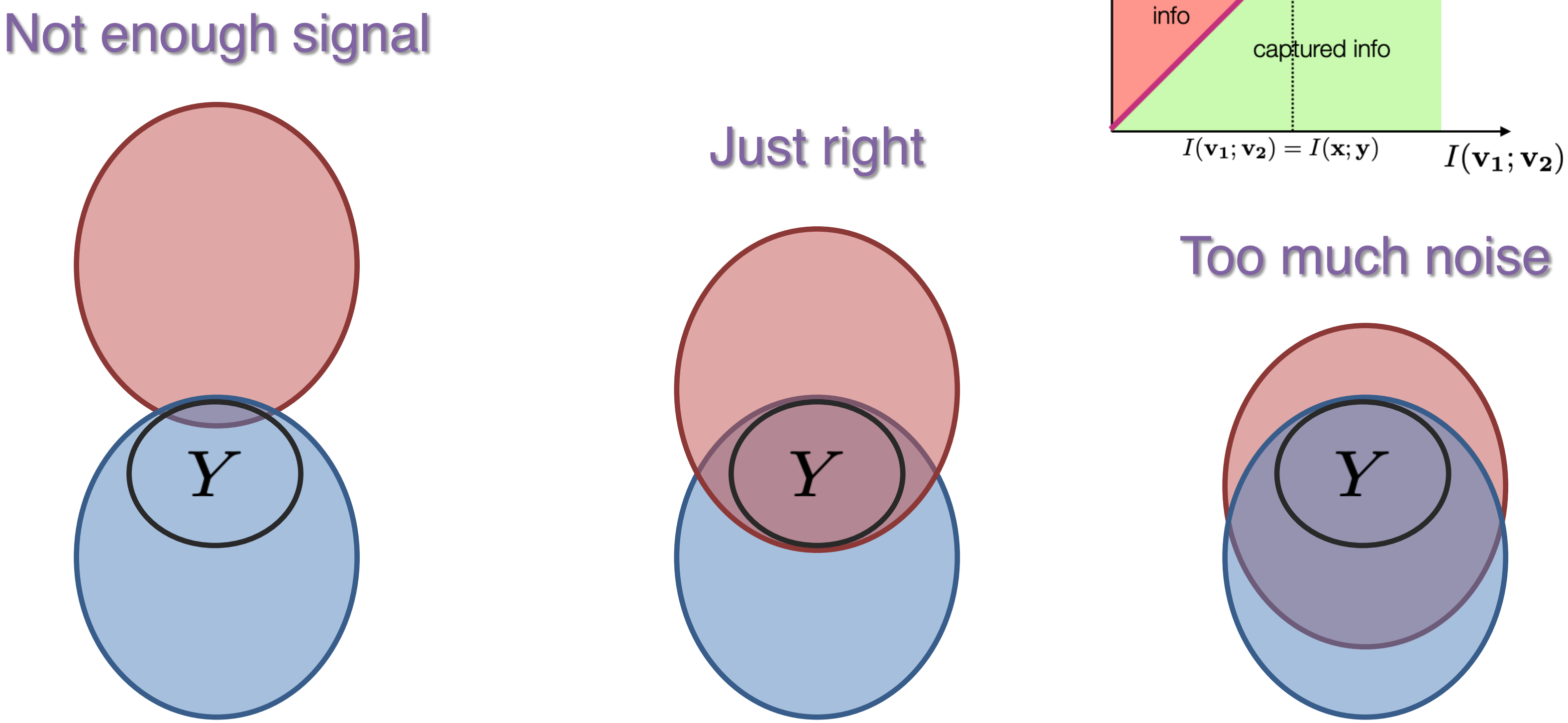
Multiview Redundancy and Contrastive Learning

Open challenges

How much information should be shared?

Multi-view redundancy: $I(X_1; X_2) = I(X_1; Y)$



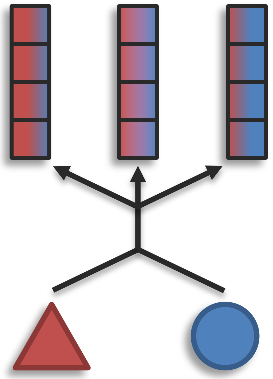


Multi-view redundancy may not hold for multimodal problems!

[Tian et al., What makes for Good Views for Contrastive Learning? NeurIPS 2020]
[Tosh et al., Contrastive Learning, Multi-view Redundancy, and Linear models. ALT 2021]

Representation Fission

Sub-Challenge 1c: Representation Fission

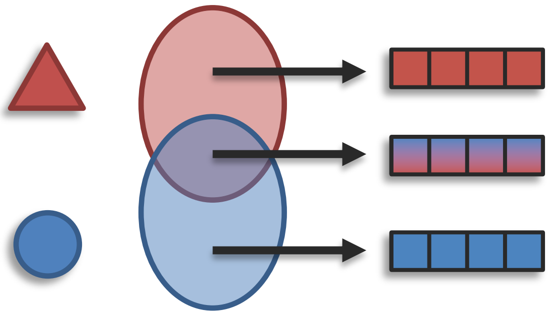


Definition: Learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

Modality-level fission:

Modality A

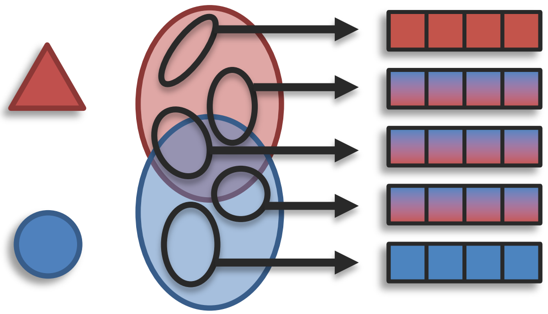
Modality B



Fine-grained fission:

Modality A

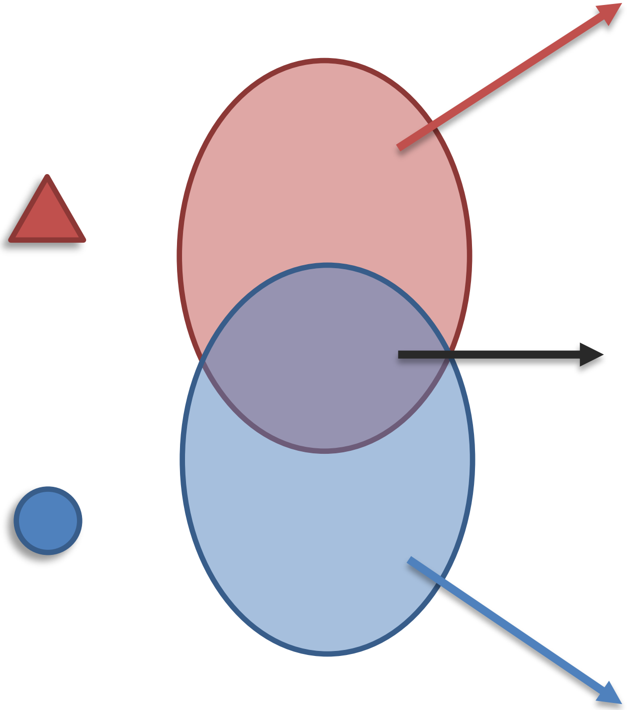
Modality B



Modality-Level Fission

Language

Visual
(image)

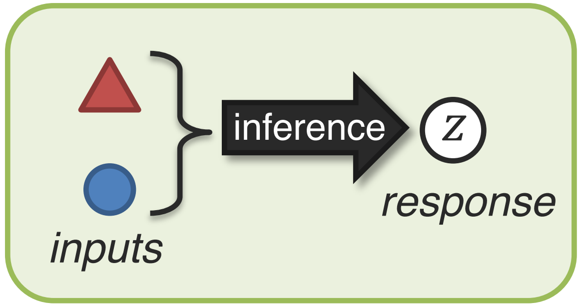


Information primarily in language modality
- Syntactic structure
- Vocabulary, morphology
- …

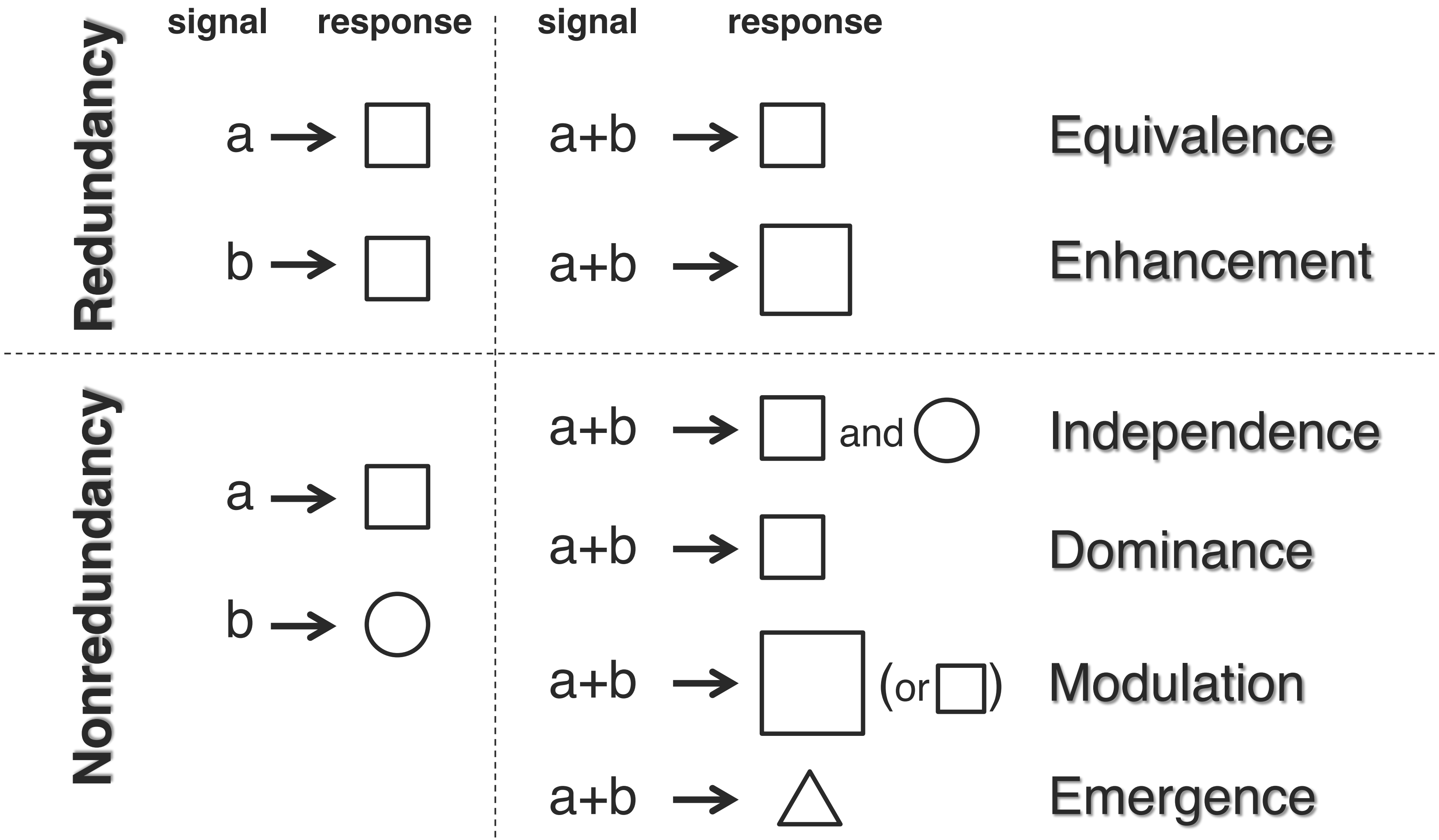
Information in both modalities
- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality
- Texture, visual appearance
- Depth, perspective, motion
- …

Recall Taxonomy of Interactions



Multimodal Communication



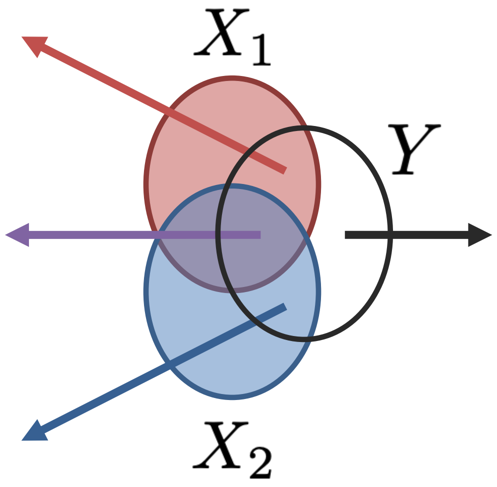
Partan and Marler (2005). *Issues in the classification of multimodal communication signals*. American Naturalist, 166(2)

Representation Fission via Information Theory

Unique to modality 1
and task Y

Redundancy: Shared by
both modalities and task

Unique to modality 2
and task Y



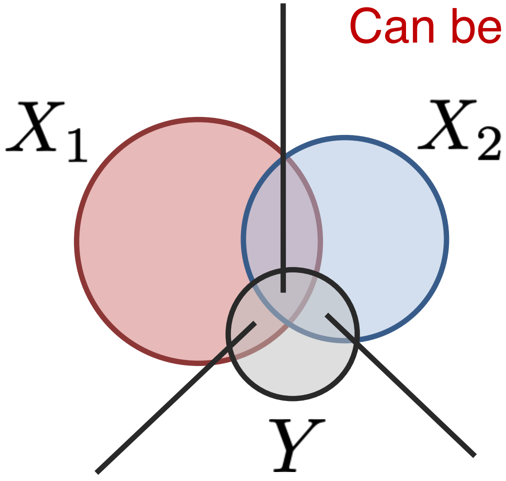
Synergy: Emerging information
from multimodal interaction

Partial Information Decomposition

Classical Information Theory

$$R = I(X_1;X_2;Y) = I(X_1;X_2) - I(X_1;X_2|Y)$$

Can be negative!

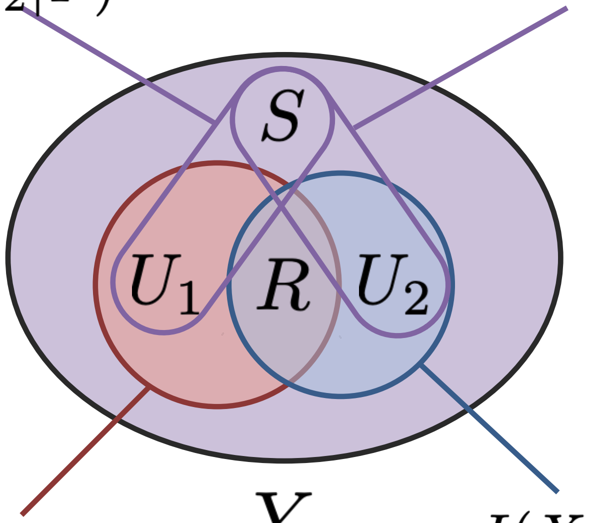


$U_1 = I(X_1;Y|X_2)$ $U_2 = I(X_2;Y|X_1)$

No synergy!

Partial Information Decomposition

$I(X_1;Y|X_2)$ $I(X_2;Y|X_1)$



$I(X_1;Y)$ Y $I(X_2;Y)$

$$R - S = I(X_1;X_2;Y)$$ Explains negative!

$$R + U_1 + U_2 + S = I(X_1,X_2;Y)$$ Task-relevant multimodal info

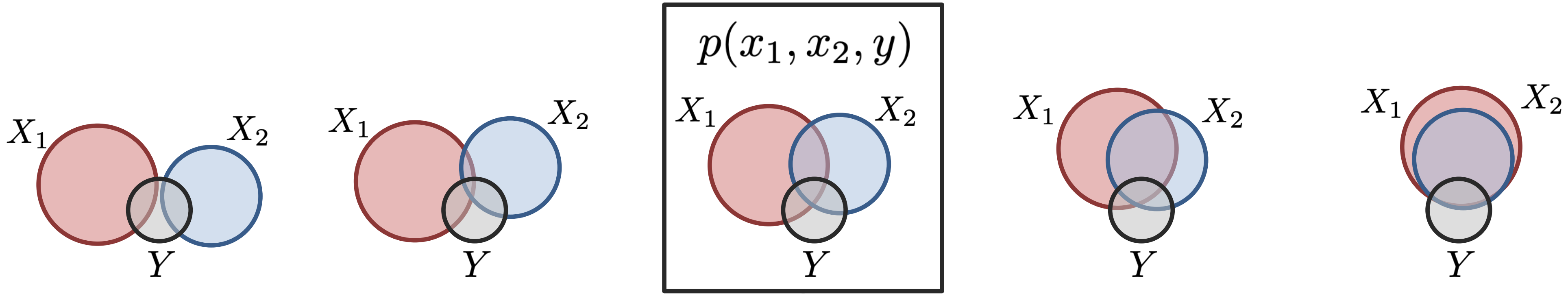
[Williams and Beer. Non-negative Decomposition of Mutual Information. 2010]

Partial Information Decomposition

One type of information decomposition

Unimodal marginal-matching distributions:

$$\Delta_p = \{q(x_1, x_2, y) : q(x_1, y) = p(x_1, y), q(x_2, y) = p(x_2, y)\}$$



$$S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$

Task-relevant multimodal info

Task-relevant multimodal info without synergy:

$$S_{q^*} = I_{q^*}(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y) = 0$$

[Bertschinger et al., Quantifying Unique Information, Entropy 2014]

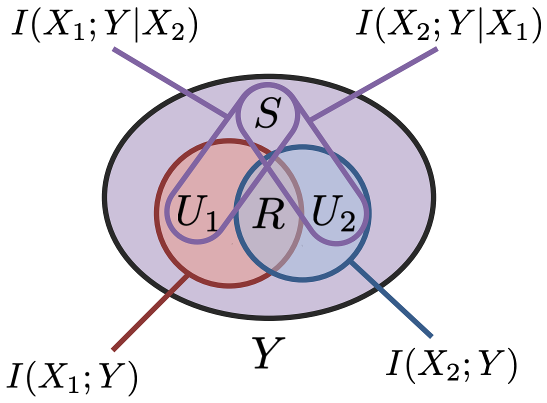
Partial Information Decomposition

One type of information decomposition

Unimodal marginal-matching distributions:

$$\Delta_p = \{q(x_1, x_2, y) : q(x_1, y) = p(x_1, y), q(x_2, y) = p(x_2, y)\}$$

$$S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$



+ consistency equations relating interactions with information theory:

Only need unimodal marginals to infer redundancy and uniqueness:

$$R = \max_{q \in \Delta_p} I_q(X_1; X_2; Y) \quad U_1 = \min_{q \in \Delta_p} I_q(X_1; Y|X_2) \quad U_2 = \min_{q \in \Delta_p} I_q(X_2; Y|X_1)$$

Can be solved efficiently as a convex optimization problem
Scales to high-dimensional continuous modalities via neural networks

[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

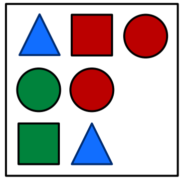
Quantifying Interactions

Open challenges

These interactions can be efficiently estimated – gives a path towards understanding interactions

Sheldon :
Its just a *privilege* to watch your mind at work.

- **Text** : suggests a compliment.
- **Audio** : neutral tone.
- **Video** : straight face.



Is there a red shape above a circle?

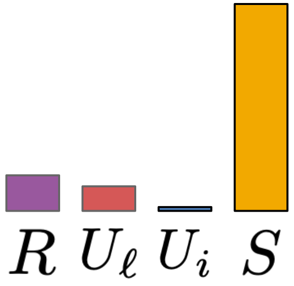
Sentiment

Language/Agreement

Sarcasm

Multimodal Transformer

VQA



Multiplicative/Transformer

Also matches human judgment of interactions, and other sanity checks on synthetic datasets

Can also be used to choose most appropriate models – can they be used to better train/design new models?

[Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. arXiv 2023]

Quantifying Interactions

Open challenges

Lower and upper bounds for interactions in a semi-supervised setting: $p(x_1, y), p(x_2, y), p(x_1, x_2)$

Idea 2: min-entropy couplings

Efficient approximation algorithms [Cicalese et al., 2002, Compton 2022]

Upper bound:

$$\overline{S} = c_2 - \min_{r \in \Delta_p} H_r(X_1, X_2, Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$

$$\overline{S} = \max_{r \in \Delta_p} I_r(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$

$$S = I_p(X_1, X_2; Y) - \min_{q \in \Delta_p} I_q(X_1, X_2; Y)$$

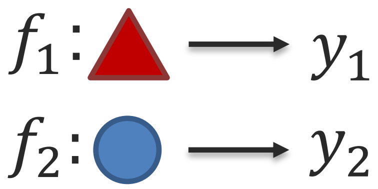
Task-relevant multimodal info

Task-relevant multimodal info without synergy

Lower bound:

$$\underline{S} = \alpha(f_1, f_2) \cdot c_1 - \max(U_1, U_2)$$

Idea 1: disagreement



Gives theoretical results on estimating interactions and model performance for semi-supervised multimodal learning

[Liang et al., Multimodal Learning Without Labeled Multimodal Data: Guarantees and Applications, arXiv 2023]

On Agreement, Disagreement, and Synergy

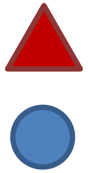
Open challenges

A (mini) taxonomy of interactions

$p(x_1, x_2, y)$

Agreement

$p(y|x_1) = p(y|x_2)$

f_1:  $\longrightarrow y_1$

f_2: $\longrightarrow y_2$

Disagreement

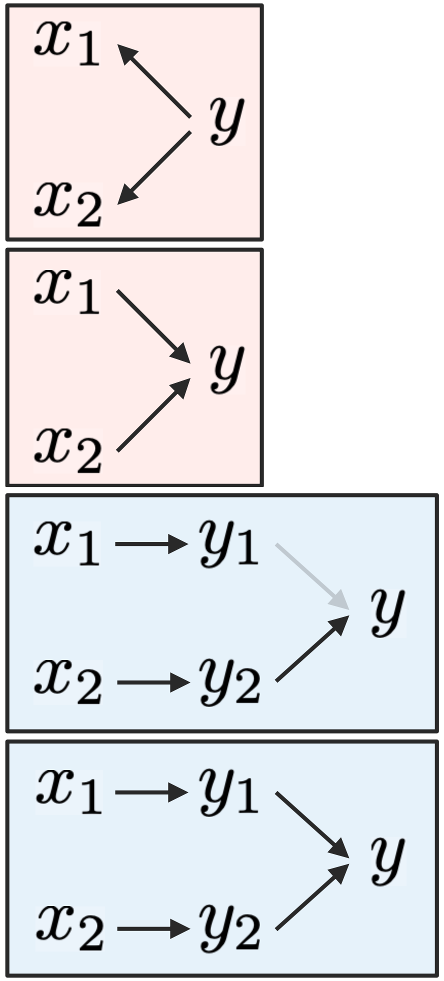
$p(y|x_1) \neq p(y|x_2)$

$R > S$

$S > R$

$U > S$

$S > U$



$I(X_1; X_2) > I(X_1; X_2|Y)$

Agreement redundancy

Contrastive learning

$I(X_1; X_2) < I(X_1; X_2|Y)$

Agreement synergy

Future work?

Disagreement uniqueness

Feature selection

Disagreement synergy

Future work?

[Blum and Mitchell. Combining Labeled and Unlabeled Data with Co-training. COLT 1998
[Peng et al., Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. TPAMI 2005]
[Liang et al., Multimodal Learning Without Labeled Multimodal Data: Guarantees and Applications, arXiv 2023]

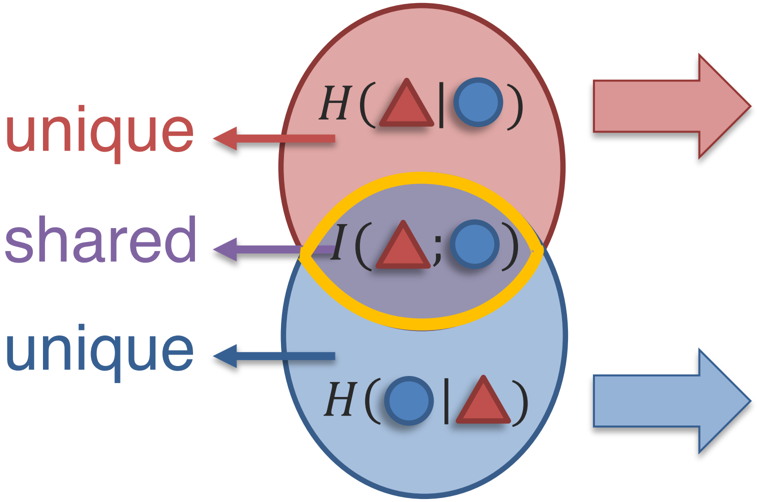
Factorized Learning of Shared + Unique Information

Modeling unique information

Can you please pass the cow?

Modality A ▲

Modality B ●



1. Maximize the mutual information

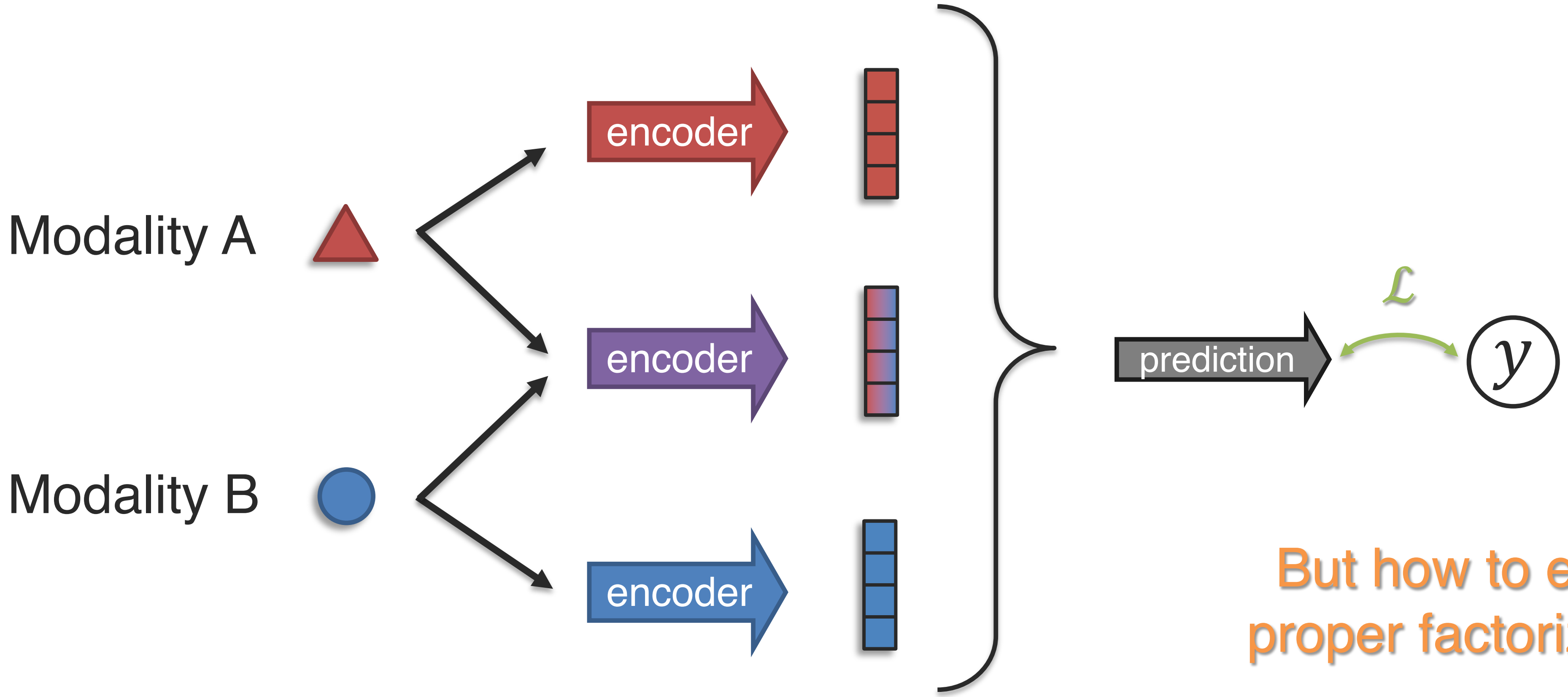
$$I(\mathbf{z};\bullet) \quad \text{and} \quad I(\mathbf{z};\blacktriangle)$$

2. Minimize the conditional entropy

$$H(\mathbf{z}|\bullet) \quad \text{and} \quad H(\mathbf{z}|\blacktriangle)$$

[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2021]
[Wang et al., Rethinking Minimal Sufficient Representation in Contrastive Learning, CVPR 2022]

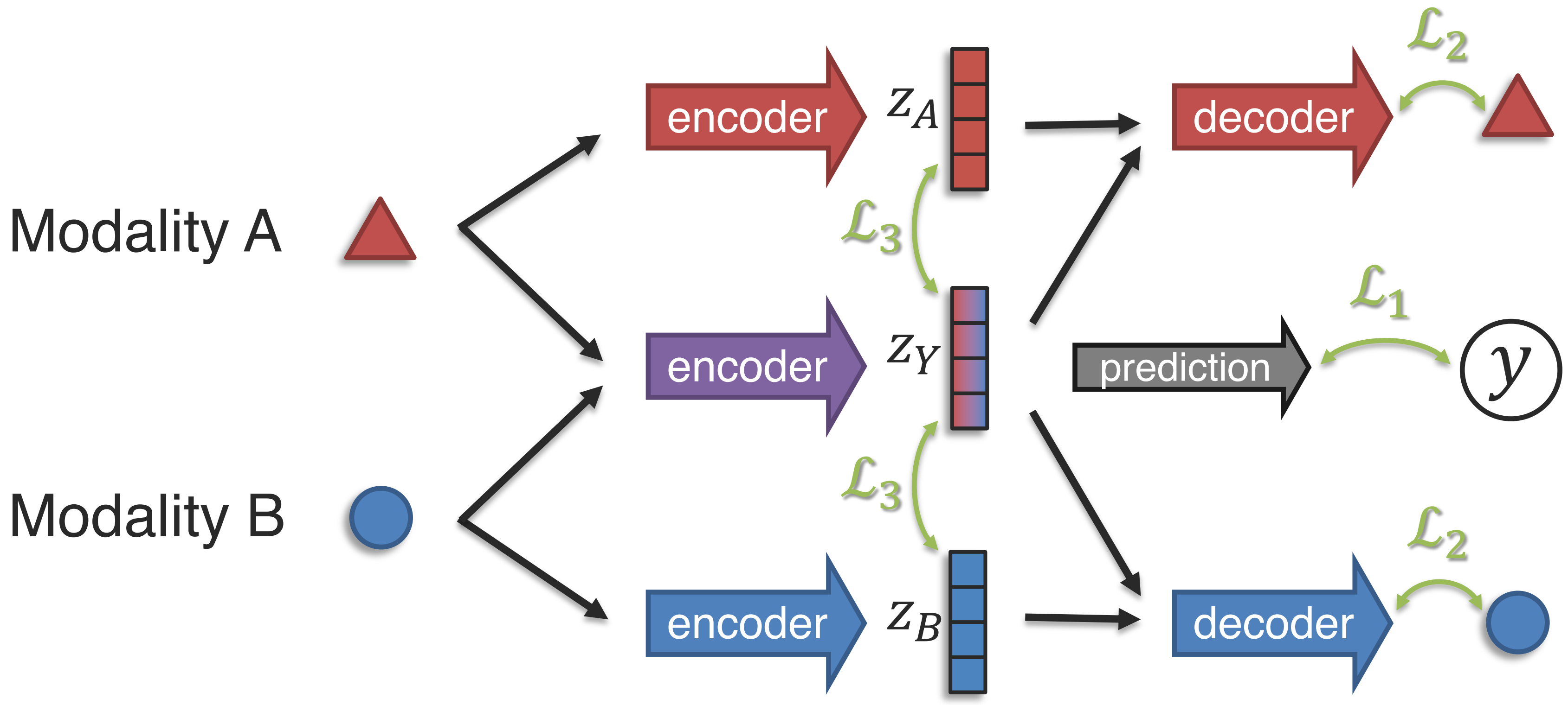
Factorized Multimodal Representations



But how to ensure proper factorization?

Tsai et al., Learning Factorized Multimodal Representations, ICLR 2019

A Generative-Discriminative Approach



$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2 + \mathcal{L}_3$

$\mathcal{L}_1$: discriminative

$\mathcal{L}_2$: generative

$\mathcal{L}_3$: no overlap

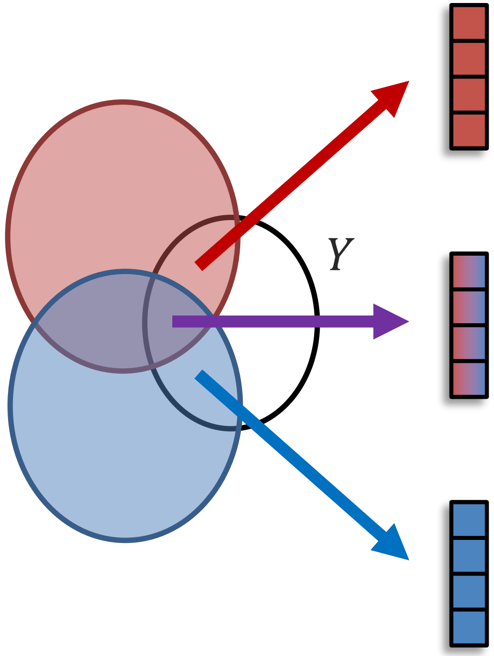
Independent priors for z_A, z_B and z_Y

Tsai et al., Learning Factorized Multimodal Representations, ICLR 2019

Learning Task-relevant Unique Information

Modeling task-relevant unique information

Can you please pass the cow?



② Maximize task-relevant **unique** information

$$I(Z; Y | \bullet)$$

① Maximize task-relevant **shared** information

$$I(Z; \bullet; Y) \quad \text{and} \quad I(Z; \blacktriangle; Y)$$

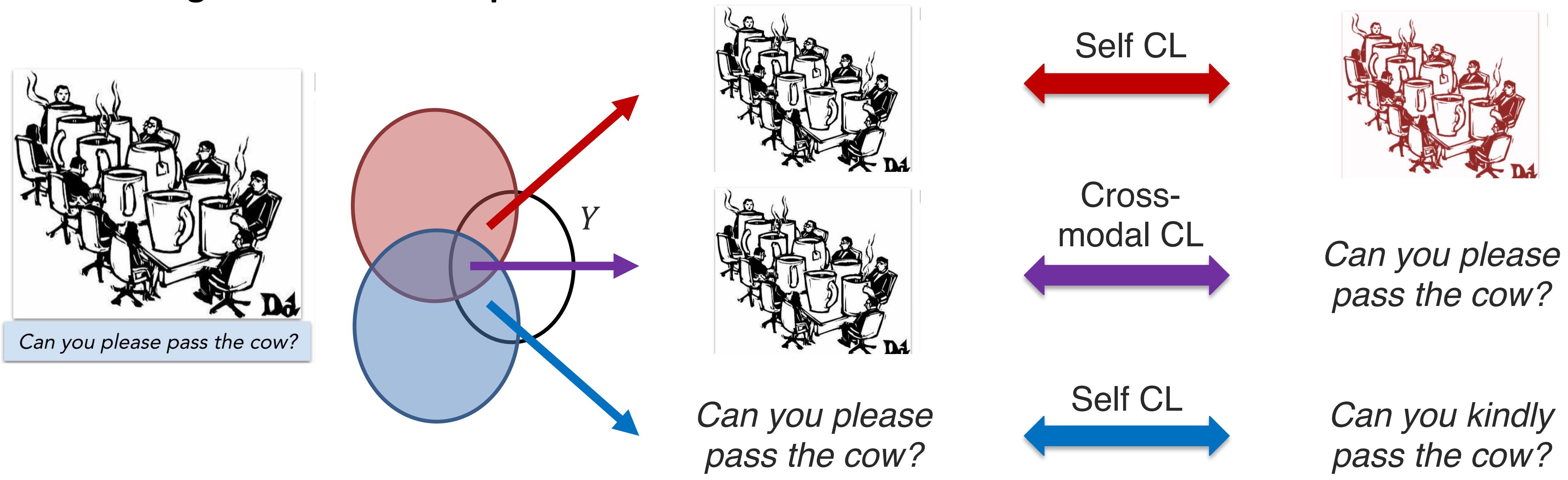
③ Maximize task-relevant **unique** information

$$I(Z; Y | \blacktriangle)$$

[Liang et al., Factorized Contrastive Learning: Going Beyond Multi-view Redundancy, arxiv 2023]

Learning Task-relevant Unique Information

Modeling task-relevant unique information



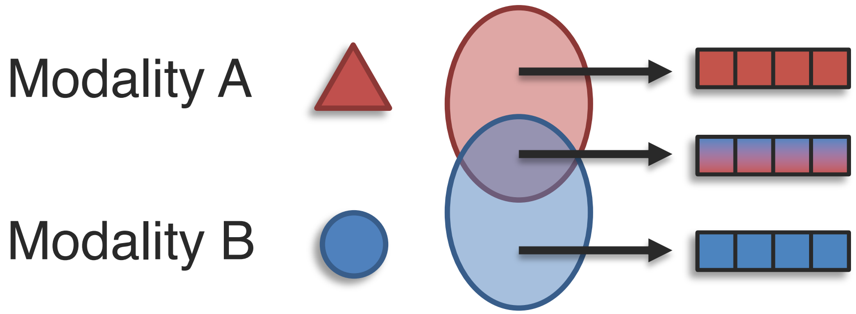
Approximate task-relevance Y using multi-view data augmentations
New scalable lower and upper bounds on mutual information

[Liang et al., Factorized Contrastive Learning: Going Beyond Multi-view Redundancy, arxiv 2023]

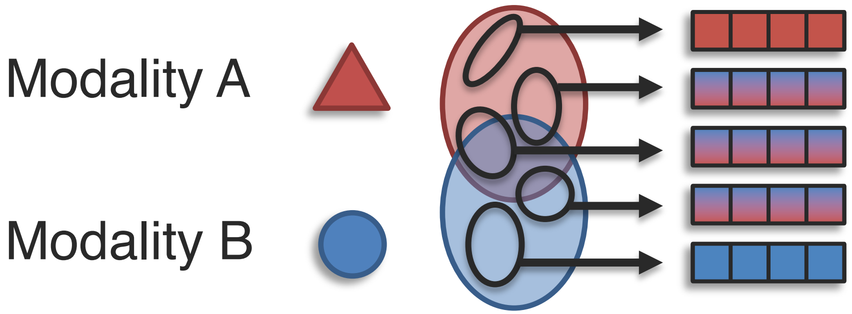
Fine-Grained Fission

How to automatically discover these internal clusters, factors?

Modality-level fission:

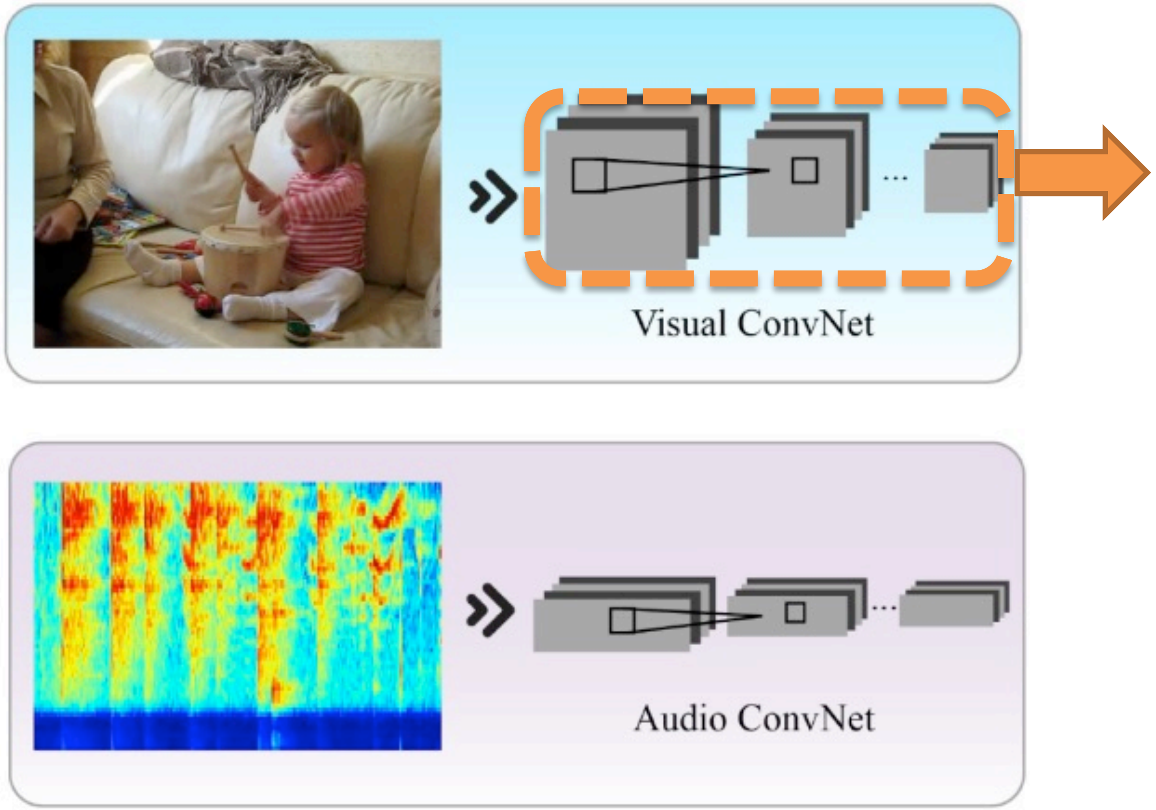


Fine-grained fission:

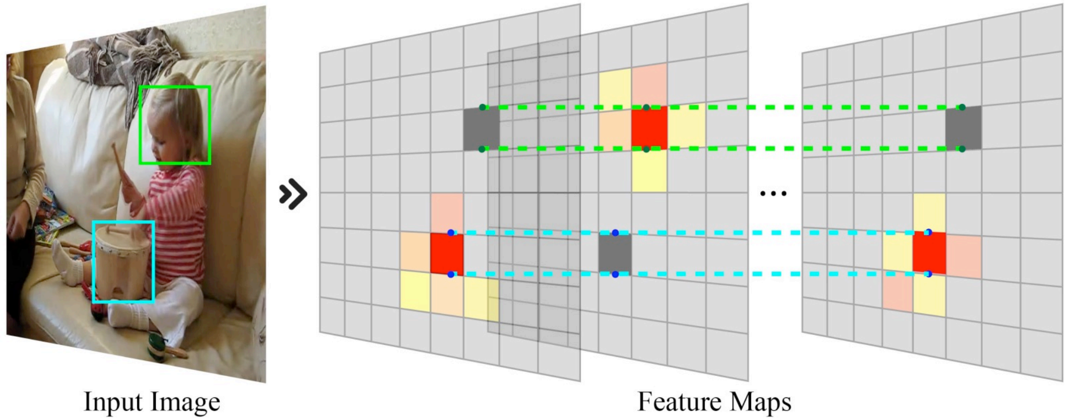


Fine-Grained Fission – A Clustering Approach

Unimodal Encoders



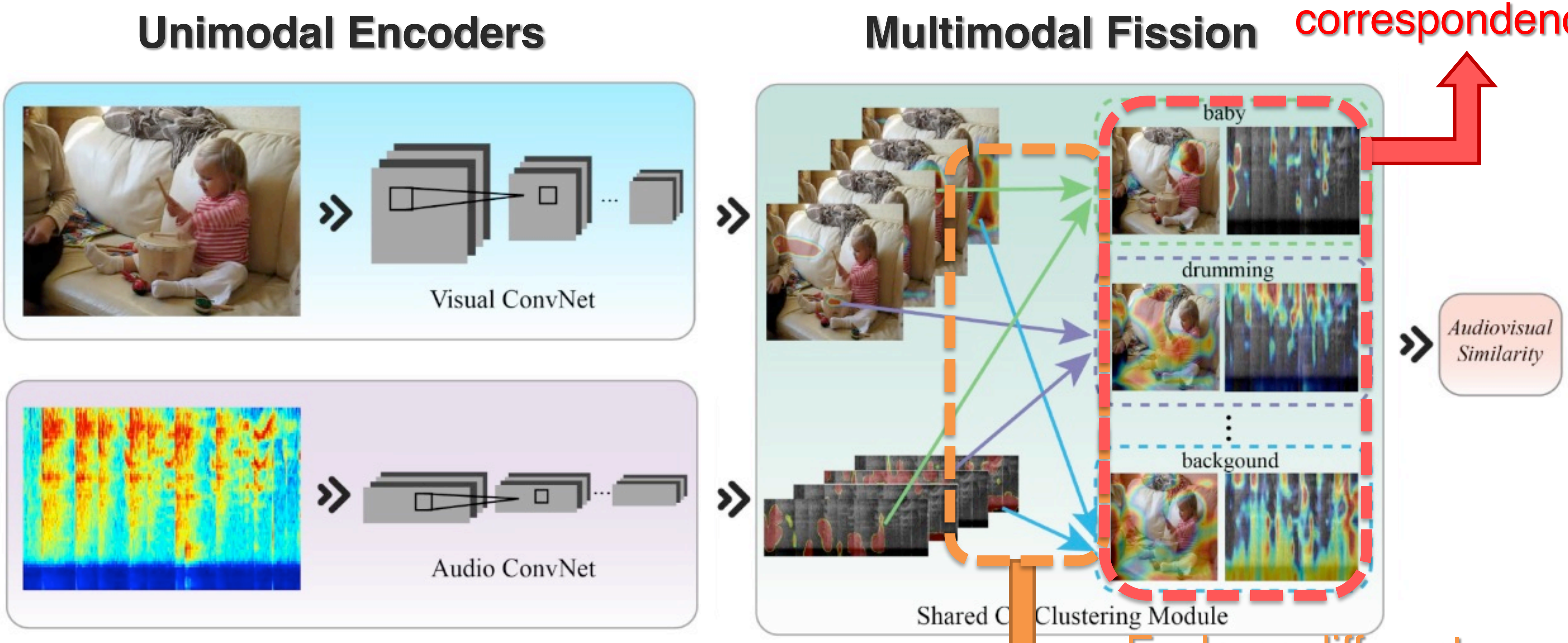
Localized activations for different objects



Hu et al., Deep Multimodal Clustering for Unsupervised Audiovisual Learning, CVPR 2019

Fine-Grained Fission – A Clustering Approach

Discovers multiple audio-visual correspondences



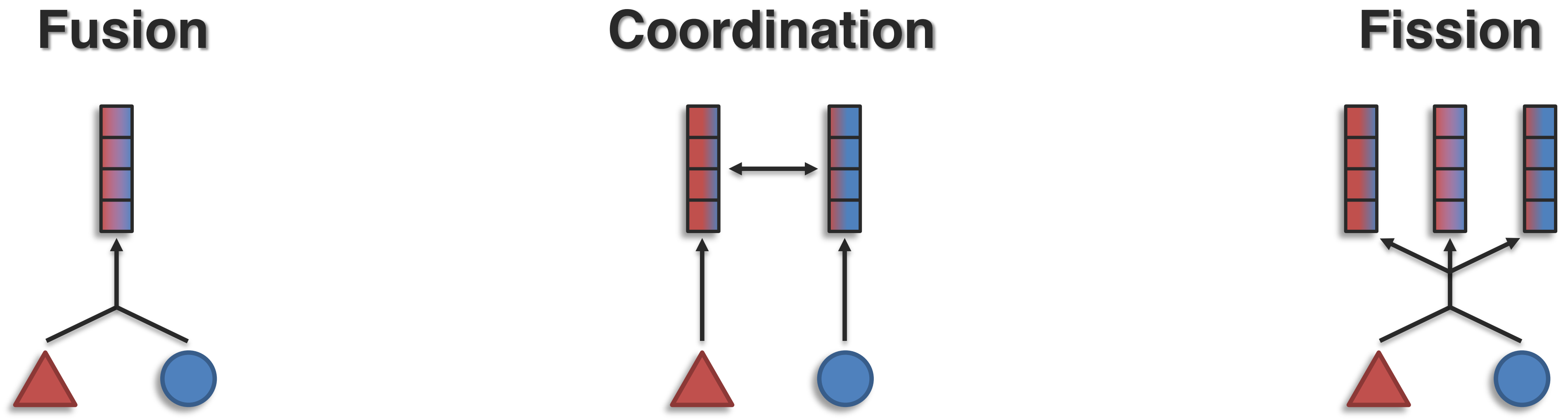
Explores different shared spaces (clusters)

Hu et al., Deep Multimodal Clustering for Unsupervised Audiovisual Learning, CVPR 2019

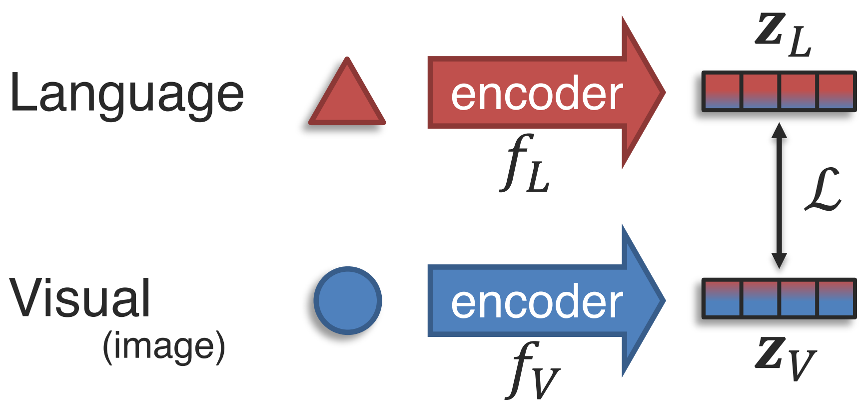
Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:



Recap: Contrastive Learning and Connected Modalities

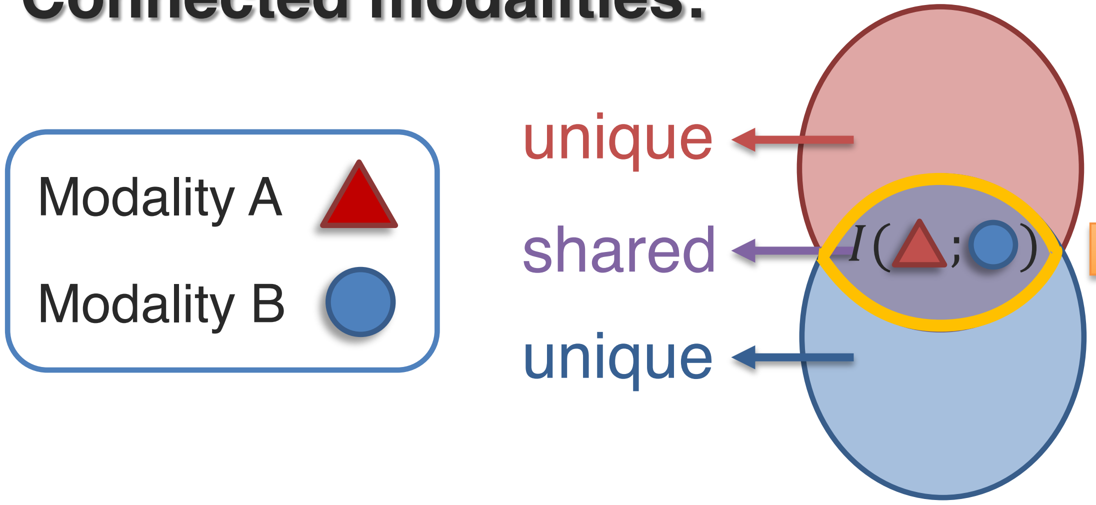


Popular contrastive loss: InfoNCE

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N}\mathrm{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

CLIP focuses on shared connections

Connected modalities:

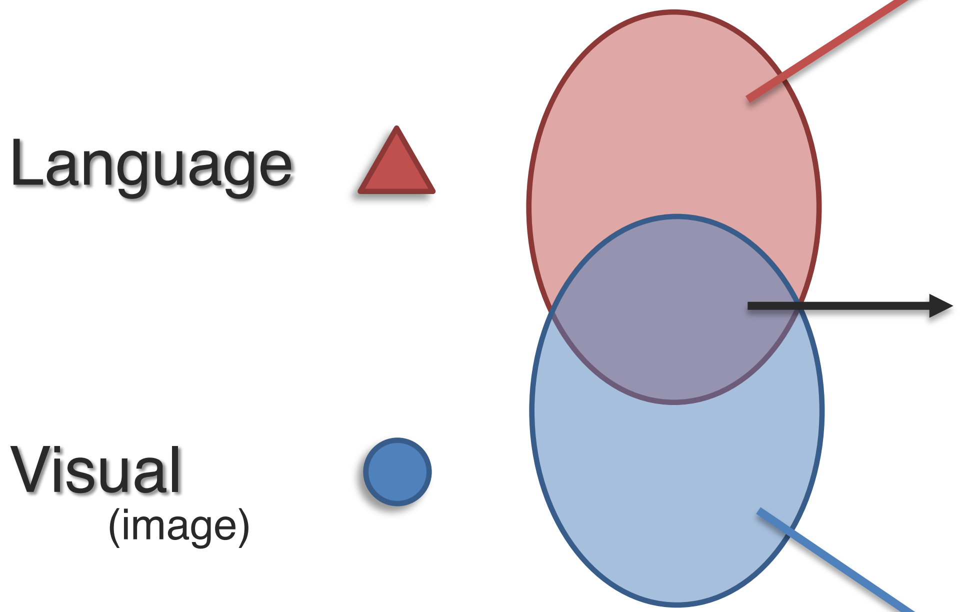


Mutual information $I(X;Y)$

$$\mathbb{E}_{X,Y}\left[\log\frac{P_{XY}(x,y)}{P_X(x)P_Y(y)}\right]$$

[Oord et al., Representation Learning with Contrastive Predictive Coding. 2018]

Recap: Modality-Level Fission



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

Information in both modalities

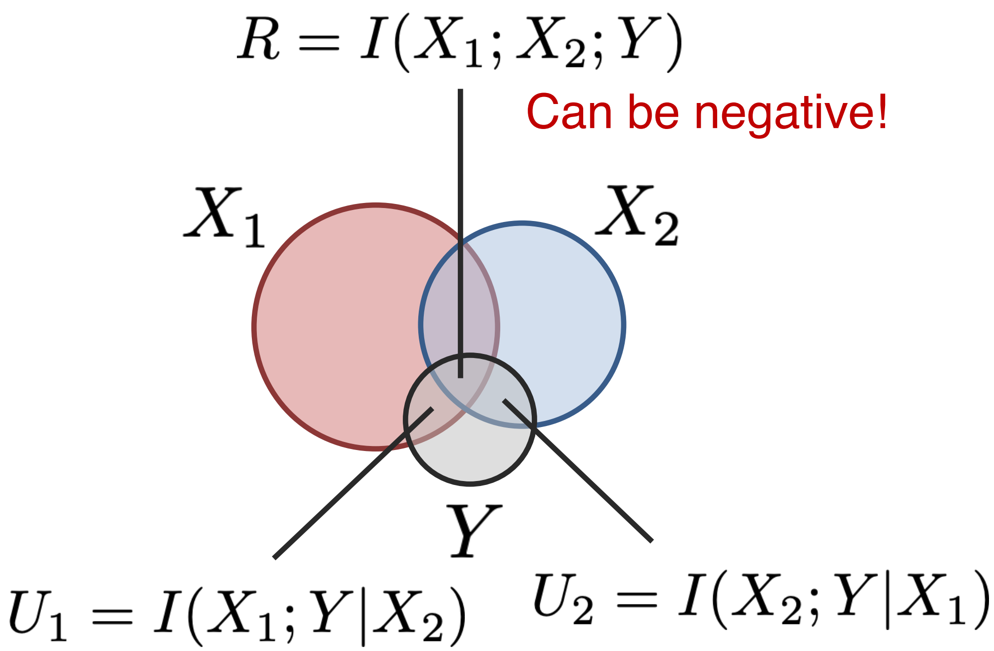
- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality

- Texture, visual appearance
- Depth, perspective, motion
- …

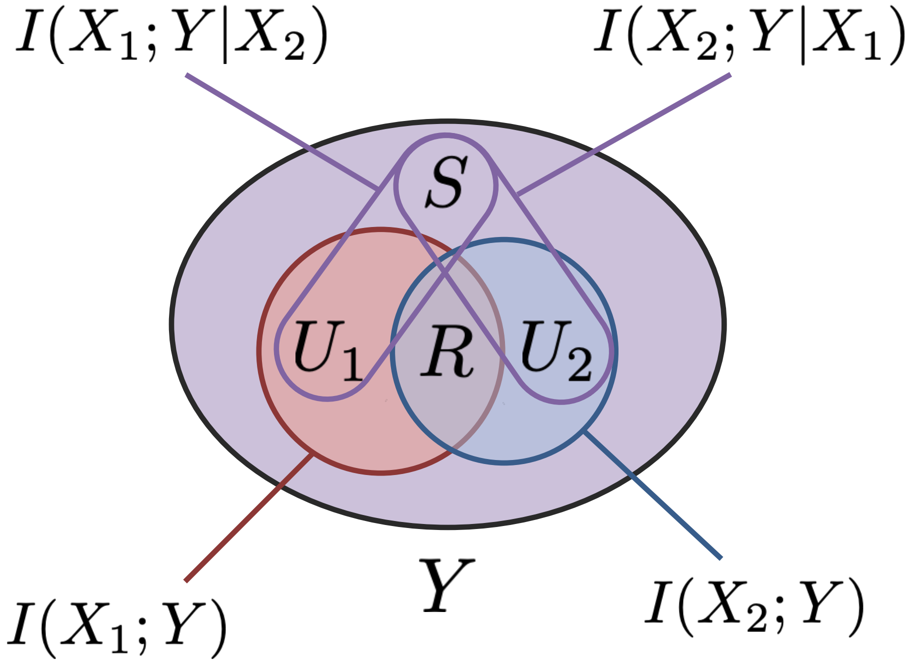
Recap: Partial Information Decomposition

Classical Information Theory



No synergy!

Partial Information Decomposition



$R - S = I(X_1; X_2; Y)$ Explains negative!

$R + U_1 + U_2 + S = I(X_1, X_2; Y)$ Task-relevant multimodal info

[Williams and Beer. Non-negative Decomposition of Mutual Information. 2010]