Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal Representations (Part 2)

Louis-Philippe Morency

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

Administrative Stuff

First Project Assignment

Due date: Sunday 9/25 at 8m

Four main sections:

- Introduction
- Related work
- Experimental setup
- Research ideas

Follows ICML paper format

The two main sections are related work and research ideas

teammates = # research ideas

Page limit depends on team size:

- 3 students : 4 pages + references
- 4 students : 4.5 pages + references
- 5 students : 5 pages + references
- 6 students : 5.5 pages + references

Team Meetings with Instructor

- No lecture on Tuesday 9/27
- 15-mins meeting with instructor
  - Optional, but highly suggested
  - Not all teammates are required to attend
  - Prepare 2 slides to summarize your research ideas
- Meetings on Tuesday 9/27 and Wednesday 9/28
- Signup form:

  https://calendly.com/morency/student-meetings

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal Representations (Part 2)

Louis-Philippe Morency

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

Objectives of today's class

- Representation fusion
  - Multimodal auto-encoder
  - Fusion from raw modalities
- Representation coordination
  - Coordination functions
    - Kernel similarity functions
    - Canonical correlation analysis
  - Contrastive learning
- Representation fission
  - Factorized multimodal representations
  - Information, entropy and mutual information
  - Clustering and fine-grained fission

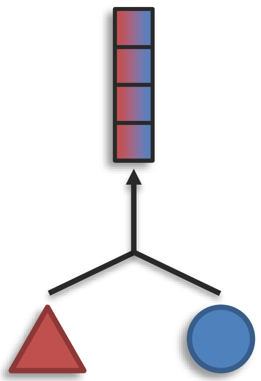
Multimodal Representation

Challenge 1: Representation

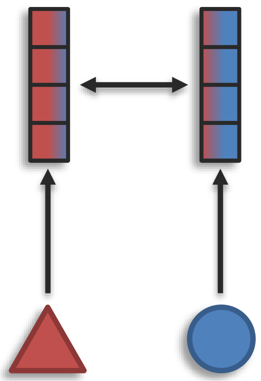
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

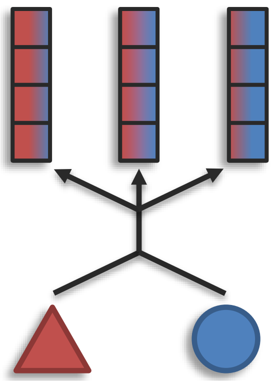
Fusion



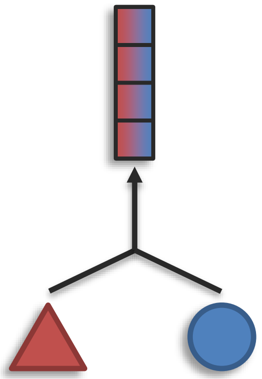
Coordination



Fission

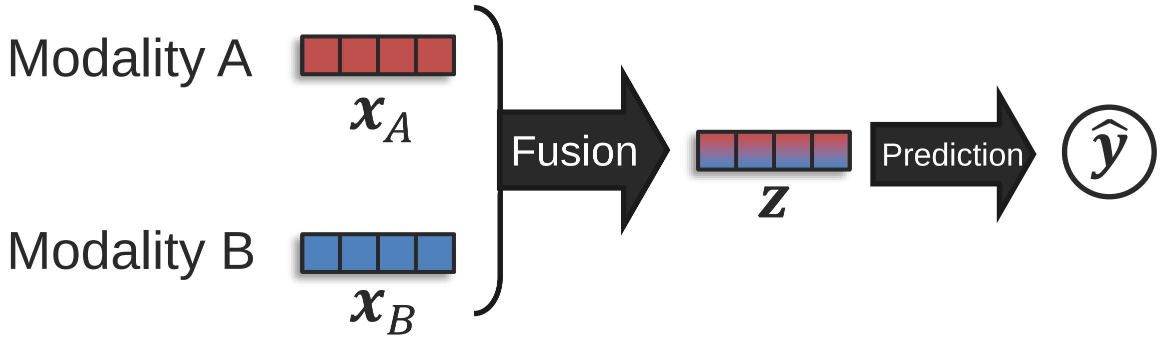


Sub-Challenge 1a: Representation Fusion



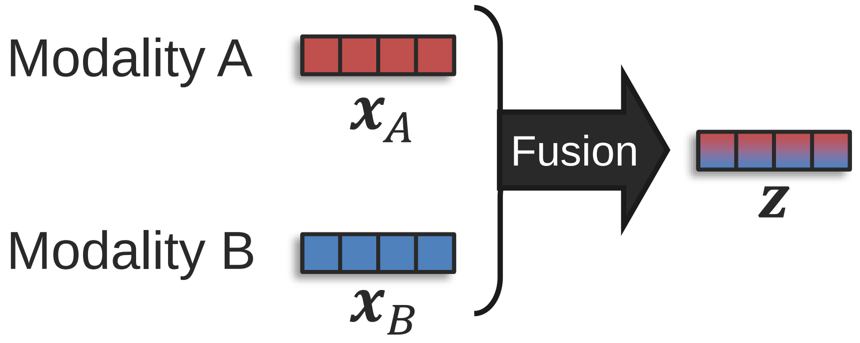
Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Learning Fusion Representations



How to learn fusion models?

Learning Fusion Representations

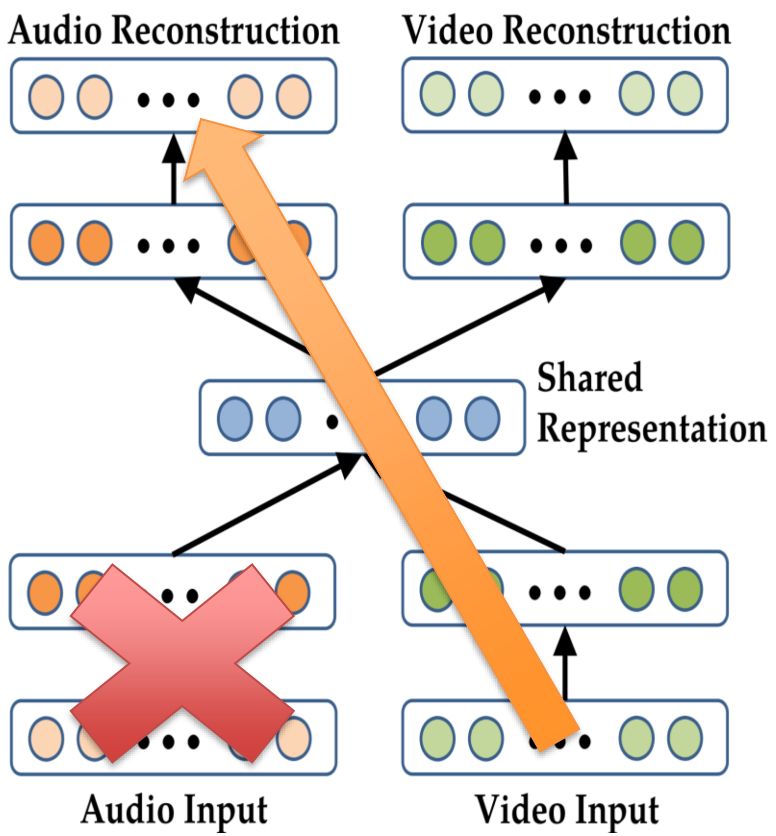


How to learn fusion models?

What will be the loss function?

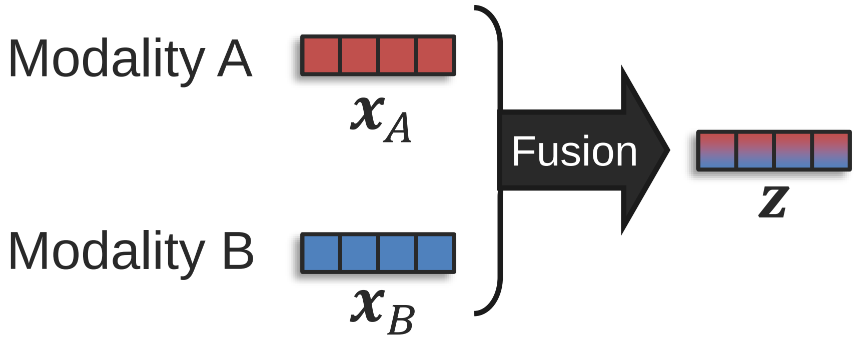
Can it hallucinate the other modality?

Multimodal Autoencoder

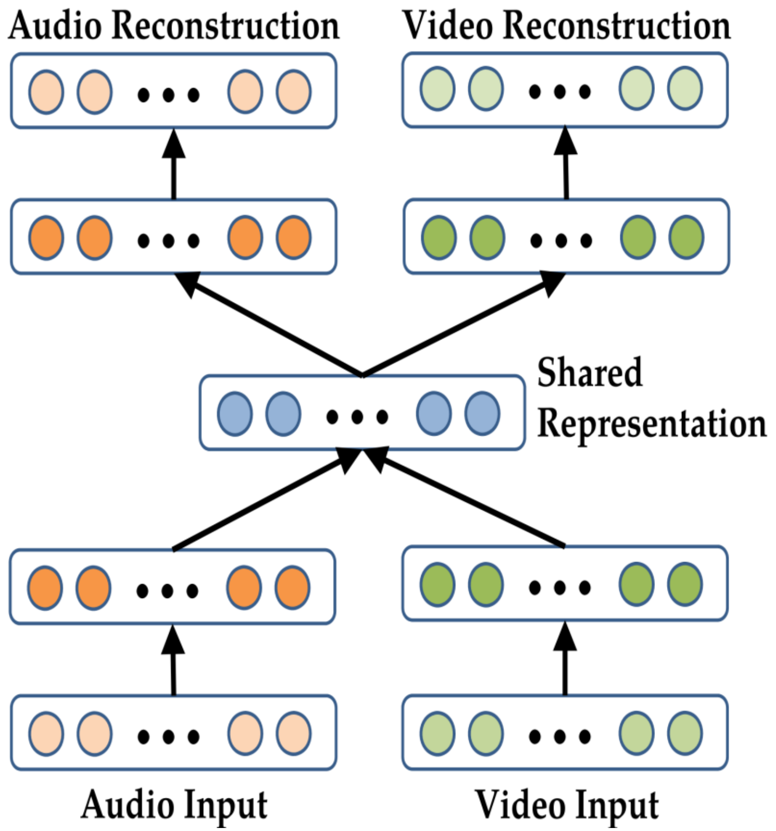


Ngiam et al, Multimodal Deep Learning, 2011

Learning Fusion Representations

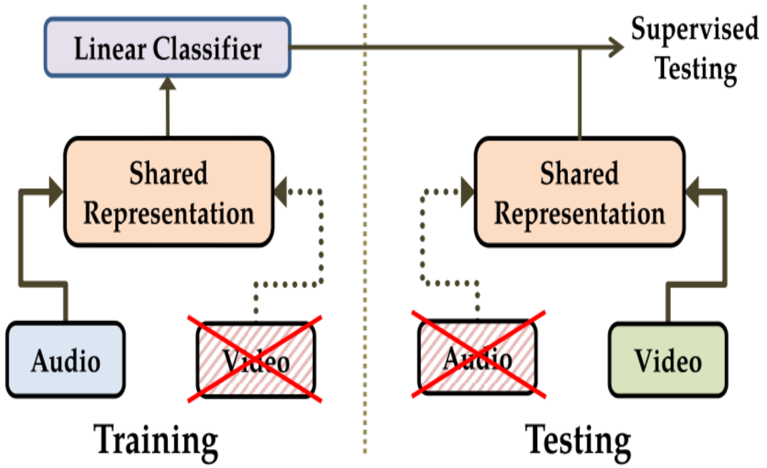


Multimodal Autoencoder



Interesting experiment: "Hearing to see"

(zero-shot cross-modal adaptation)



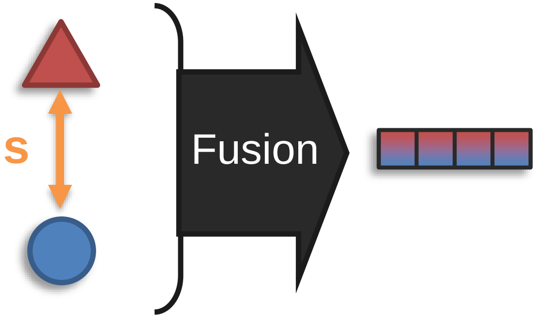
Ngiam et al, Multimodal Deep Learning, 2011

Fusion with Raw Modalities

Modality A

Heterogeneous

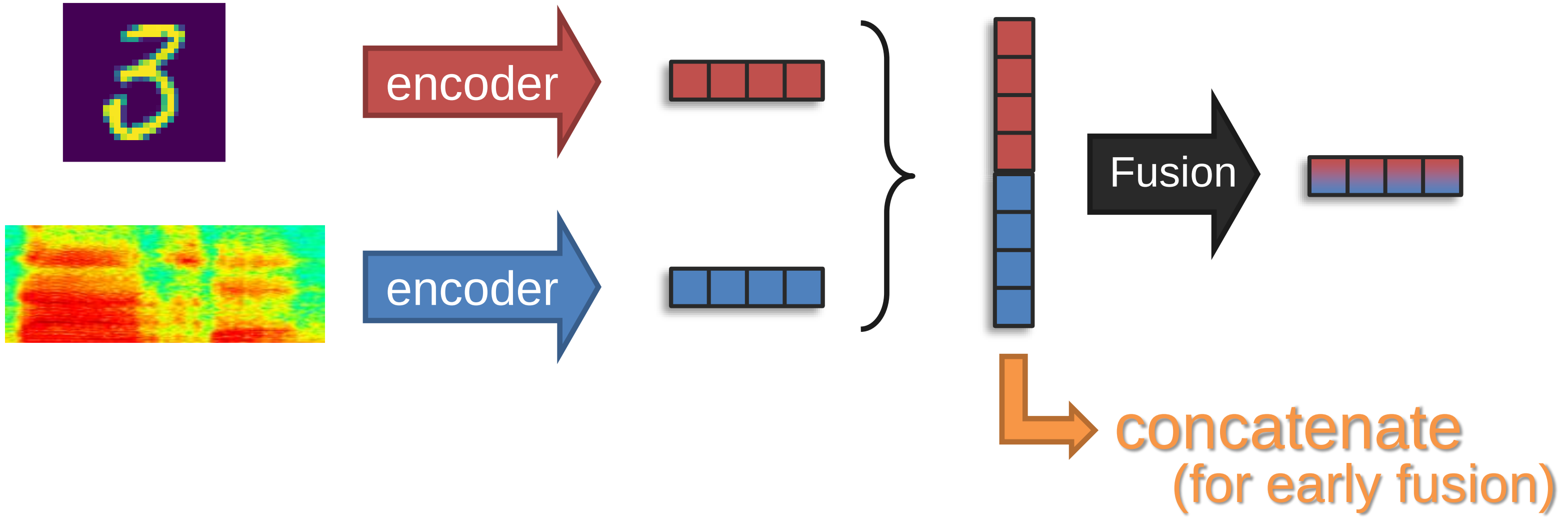
Modality B



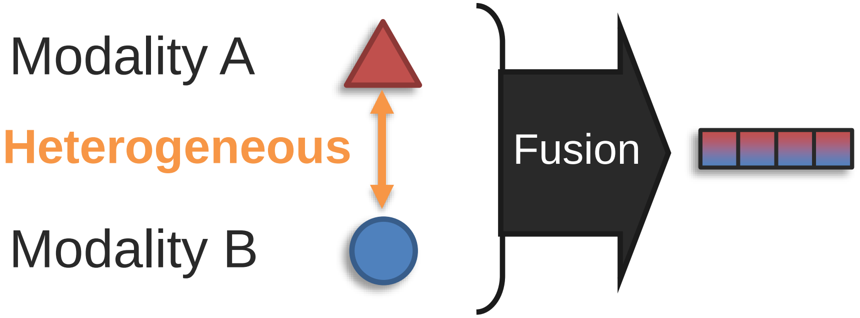
Example: From Early Fusion…

Visual

Acoustic

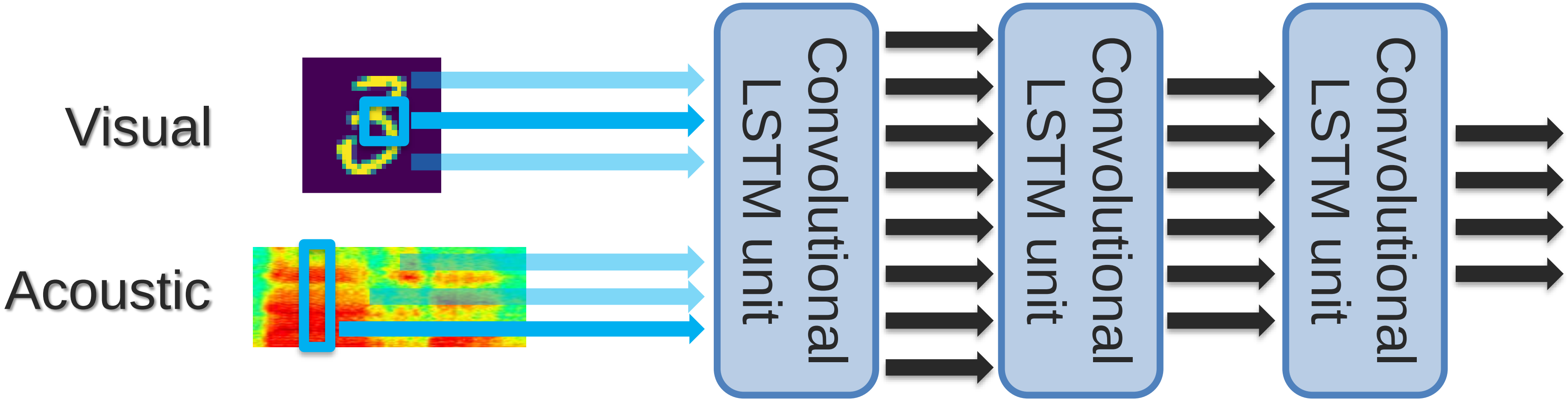


Fusion with Raw Modalities



Open Challenge!

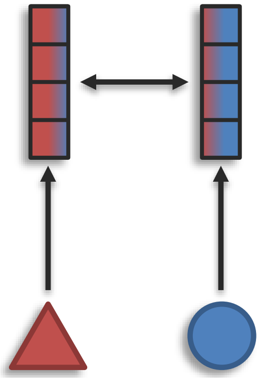
Example: From Early Fusion… to Very Early Fusion (inspired by human brain)



Barnum, et al. "On the Benefits of Early Fusion in Multimodal Representation Learning." *arxiv 2022*

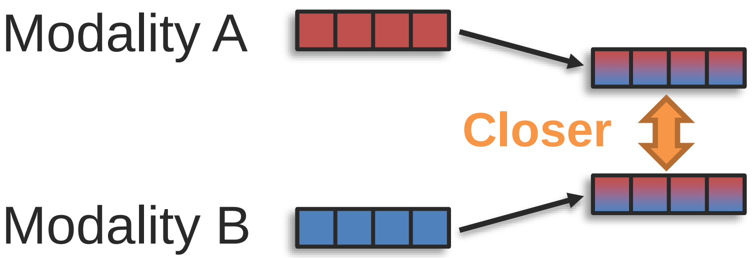
Representation Coordination

Sub-Challenge 1b: Representation Coordination

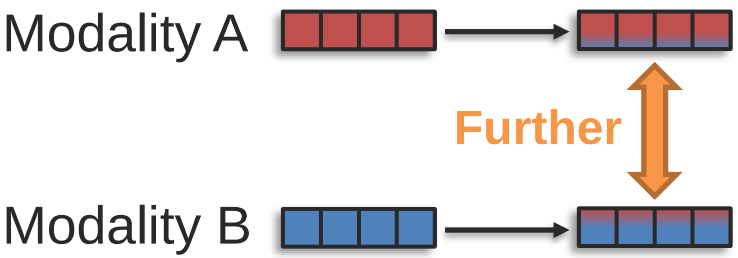


Definition: Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions

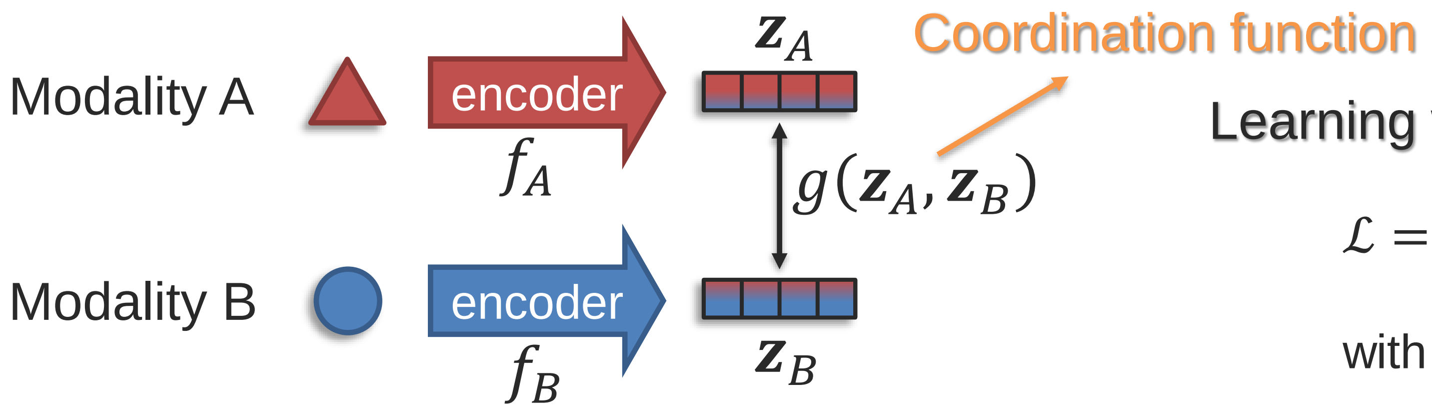
Strong Coordination:



Partial Coordination:



Coordination Function



Learning with coordination function:

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

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

➡ Requires paired data

Examples of coordination function:

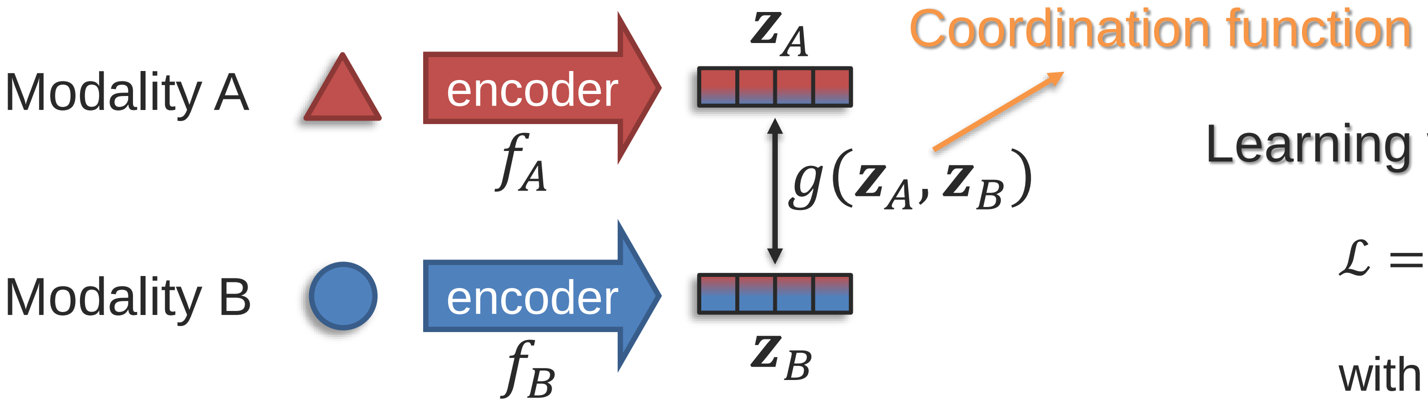
① Cosine similarity:

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

Strong coordination!

➡ For normalized inputs (e.g., $\boldsymbol{z}_A - \overline{\boldsymbol{z}_A}$) , 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

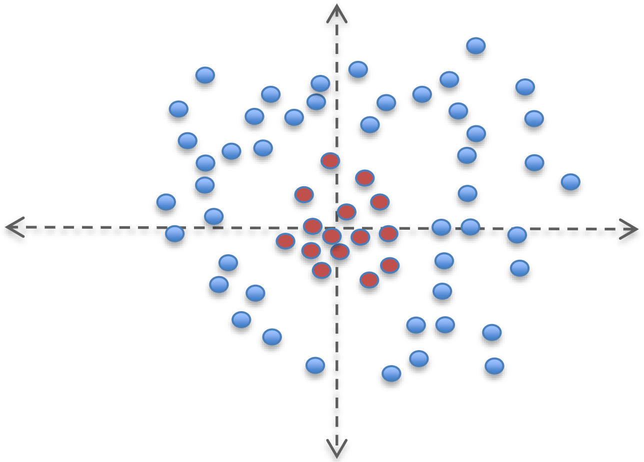
All these examples bring relatively strong coordination between modalities

Kernel Function

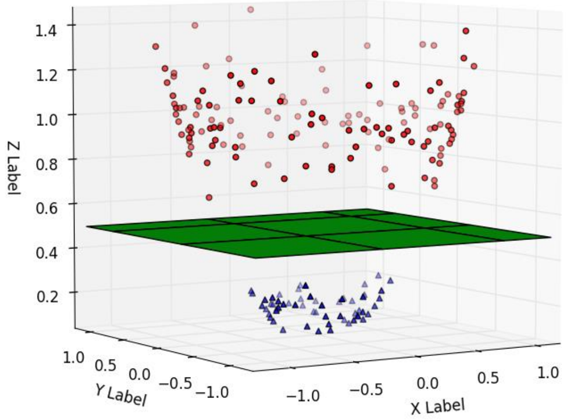
A kernel function: Acts as a similarity metric between data points

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

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



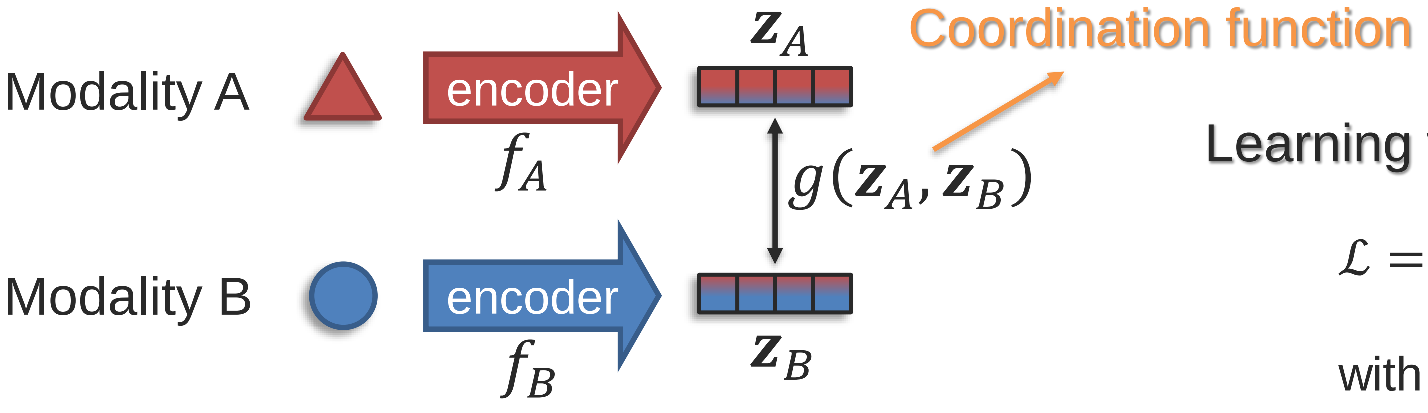
Not linearly separable in x space



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

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

Coordination Function



Learning with coordination function:

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

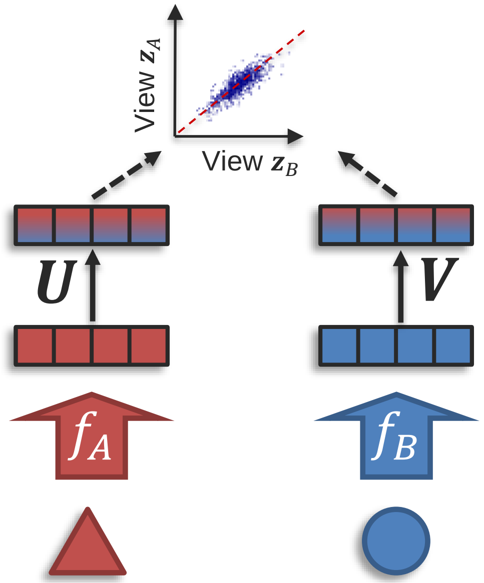
with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

③ Canonical Correlation Analysis (CCA):

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

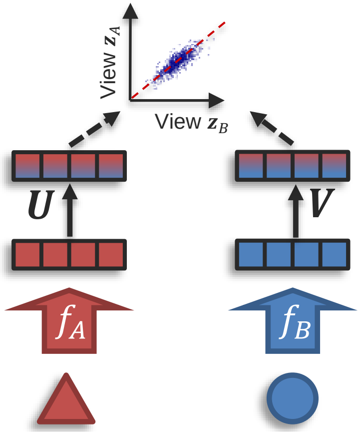
➡ CCA includes multiple projections, all orthogonal with each others

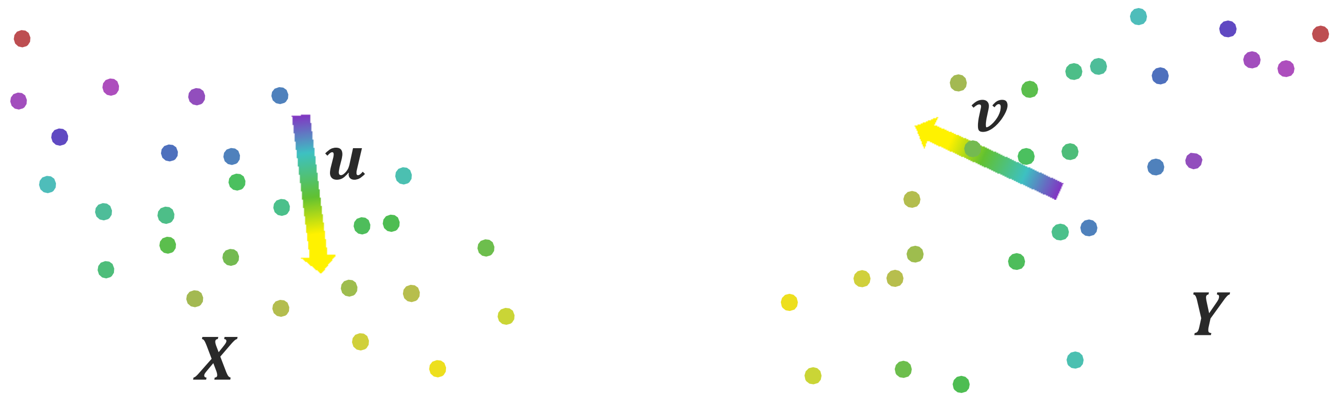


Correlated Projection

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

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



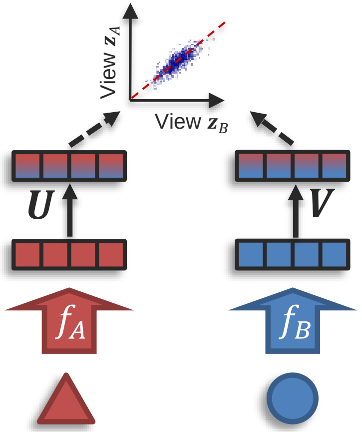


Two views $\boldsymbol{X}, \boldsymbol{Y}$ where same instances have the same color

Remember that X and Y consist of paired data

Canonical Correlation Analysis

2 We want these multiple projection pairs to be orthogonal ("canonical") to each other:



$$\boldsymbol{u}_{(i)}^{T}\boldsymbol{\Sigma}_{XY}\boldsymbol{v}_{(j)} = \boldsymbol{u}_{(j)}^{T}\boldsymbol{\Sigma}_{XY}\boldsymbol{v}_{(i)} = \boldsymbol{0} \qquad \text{for } i \neq j$$

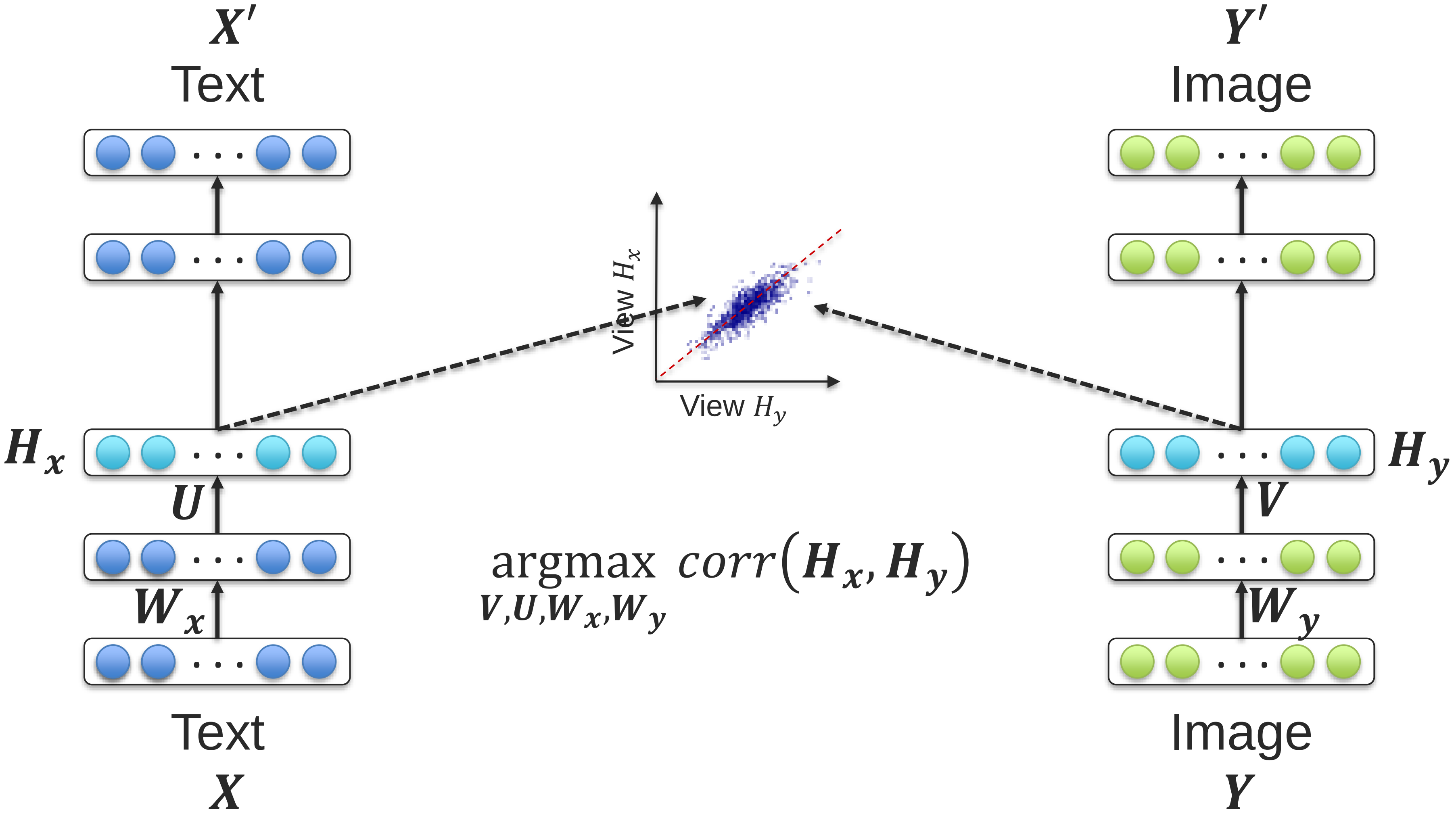
$$|\boldsymbol{U}\boldsymbol{\Sigma}_{XY}\boldsymbol{V}| = tr(\boldsymbol{U}\boldsymbol{\Sigma}_{XY}\boldsymbol{V}) \qquad \text{where } \boldsymbol{U} = [\boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \ldots, \boldsymbol{u}_{(k)}]$$

$$\text{and } \boldsymbol{V} = [\boldsymbol{v}_{(1)}, \boldsymbol{v}_{(2)}, \ldots, \boldsymbol{v}_{(k)}]$$

3 Since this objective function is invariant to scaling, we can constraint the projections to have unit variance:

$$\boldsymbol{U}^{T}\boldsymbol{\Sigma}_{XX}\boldsymbol{U} = \boldsymbol{I} \qquad \boldsymbol{V}^{T}\boldsymbol{\Sigma}_{YY}\boldsymbol{V} = \boldsymbol{I}$$

Deep Canonically Correlated Autoencoders (DCCAE)

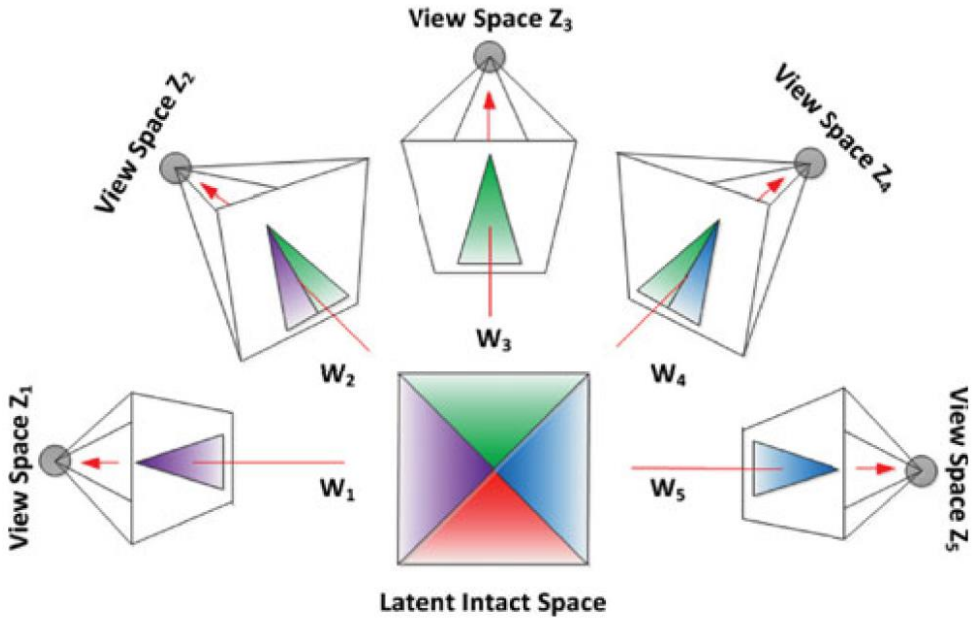


$$\underset{V,U,W_x,W_y}{\operatorname{argmax}}\ corr(H_x, H_y)$$

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":

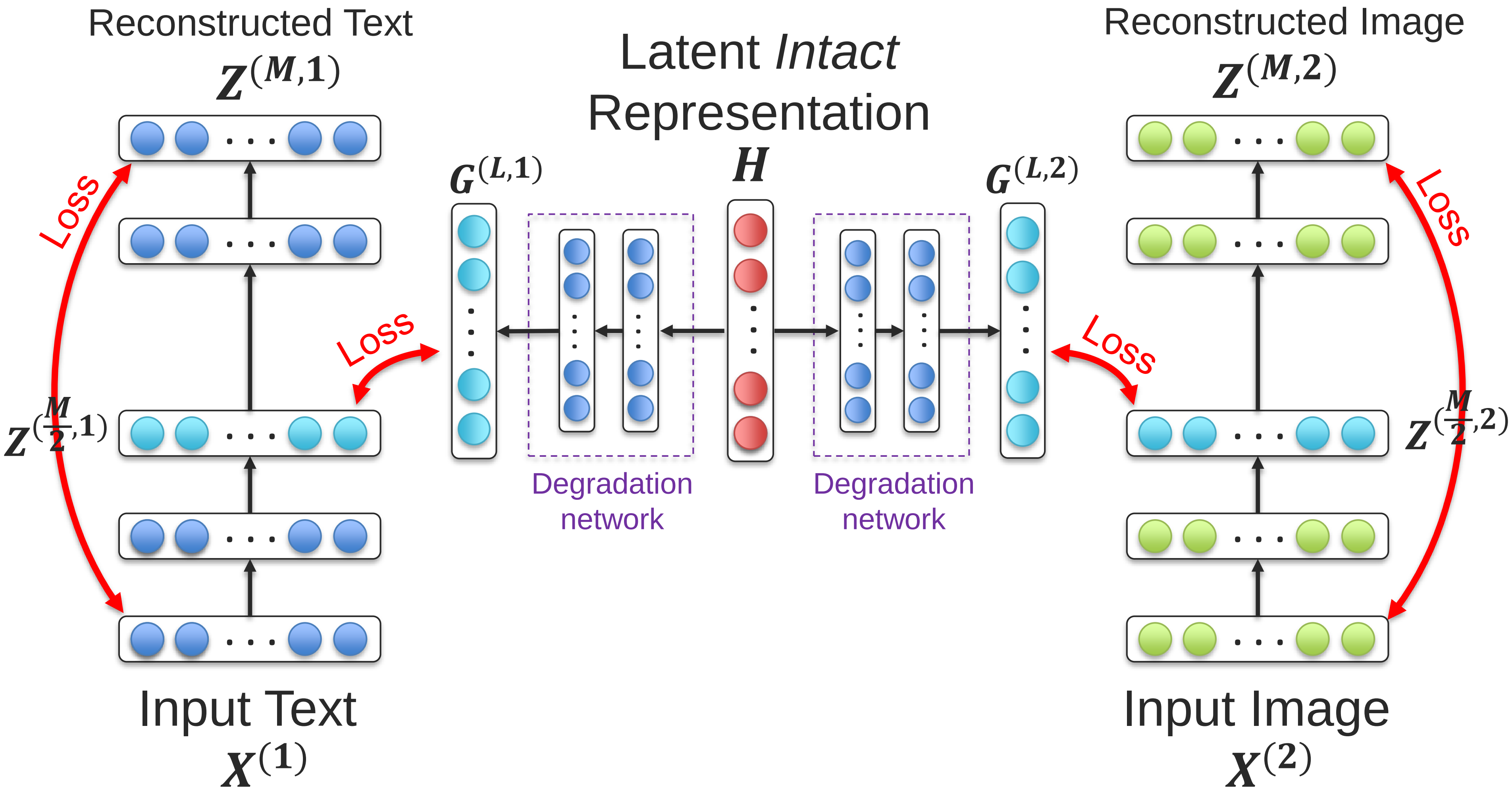


1) There is an "intact" representation which is *complete* and *not damaged*

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

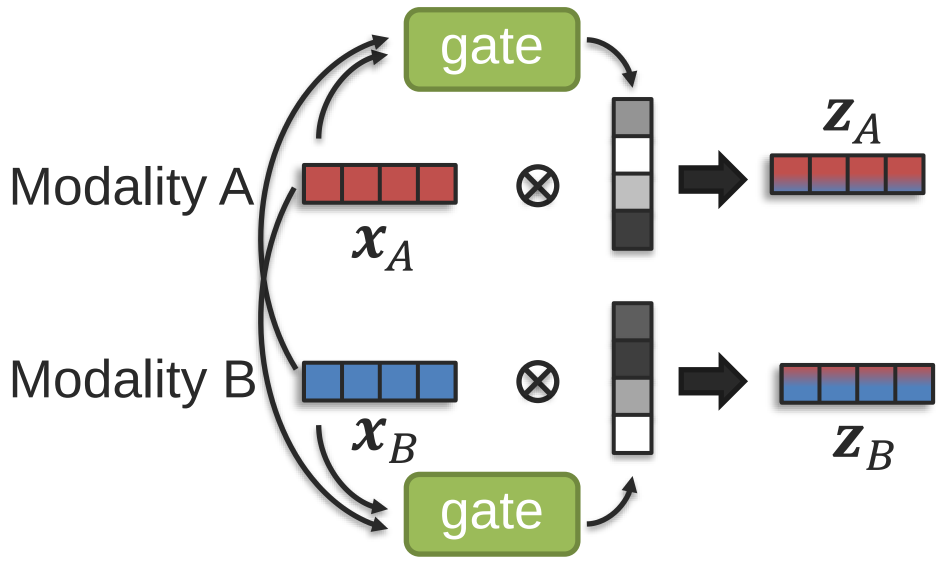
Xu et al., Multi-View Intact Space Learning, TPAMI 2015

Auto-Encoder in Auto-Encoder Network



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

Gated Coordination



Gated coordination:

$$\boldsymbol{z}_A = g_A(\boldsymbol{x}_A, \boldsymbol{x}_B) \cdot \boldsymbol{x}_A$$

$$\boldsymbol{z}_B = g_B(\boldsymbol{x}_A, \boldsymbol{x}_B) \cdot \boldsymbol{x}_B$$

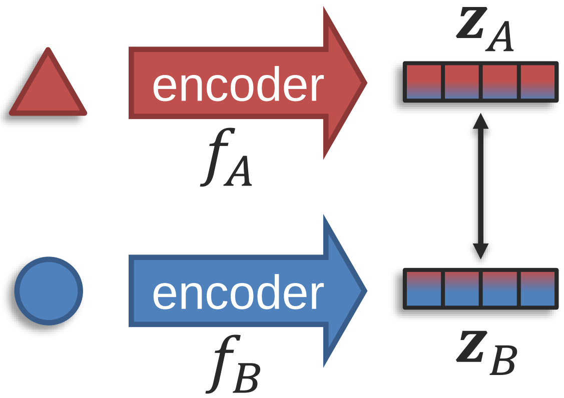
➡ Related to attention modules in transformers

More about it next week!

Coordination with Contrastive Learning

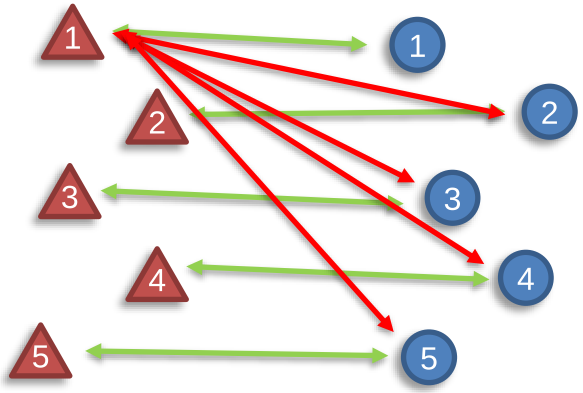
Modality A

Modality B



Paired data: { , }

(e.g., images and text descriptions)



Positive pairs

Negative pairs

Contrastive loss:

brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

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

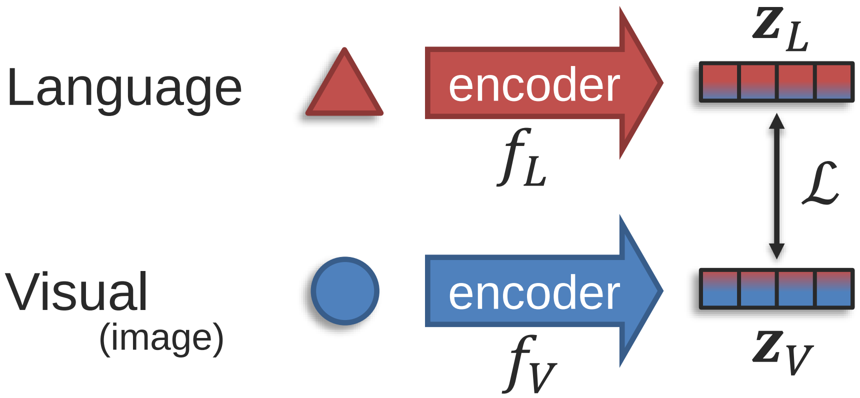
positive pairs

negative pair

Similarity functions are often cosine similarity

Similar to hinge loss

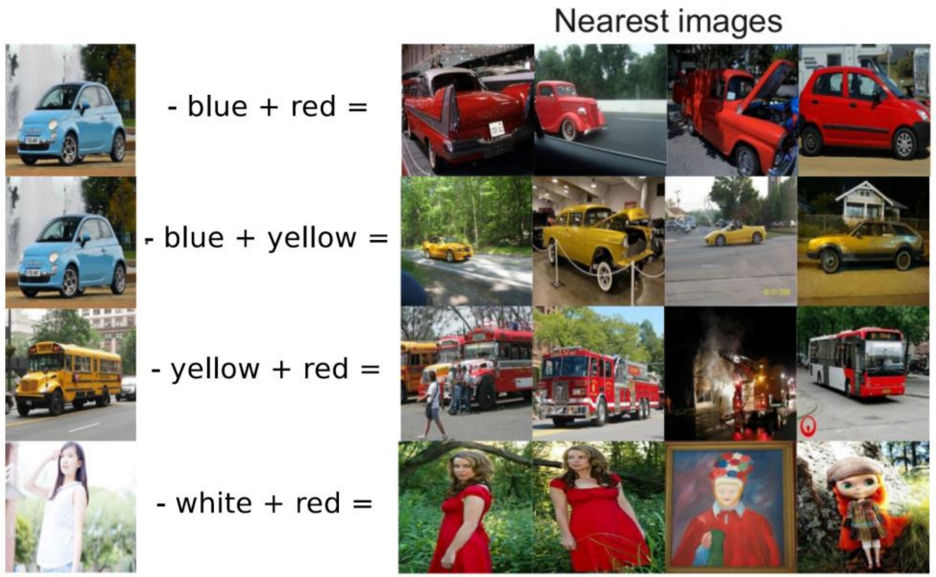
Example – Visual-Semantic Embeddings

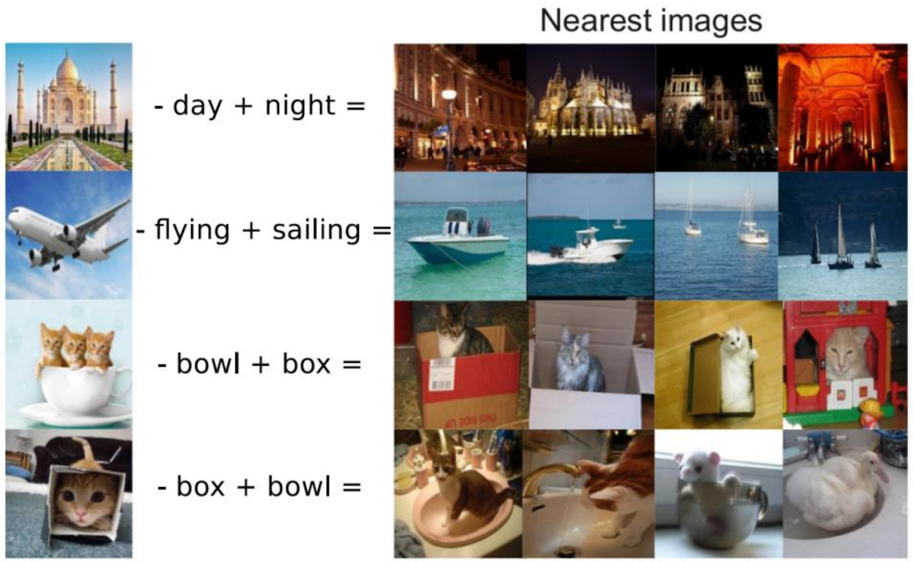


Two contrastive loss terms:

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

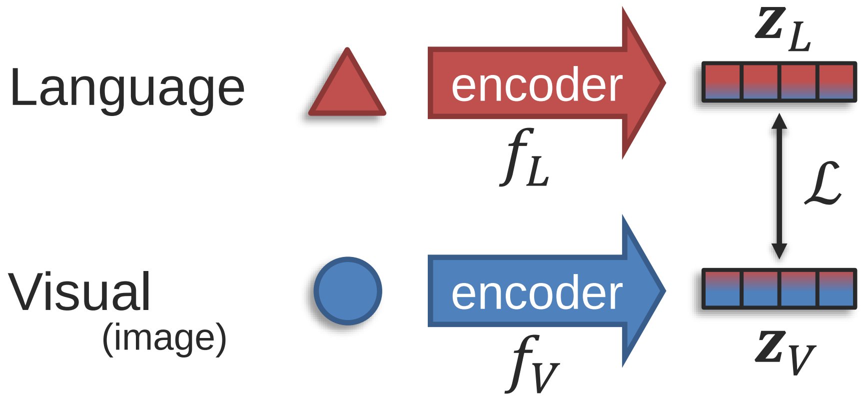
$$+ \max\{0, \alpha + sim(\boldsymbol{z}_V, \boldsymbol{z}_L^+) - sim(\boldsymbol{z}_V, \boldsymbol{z}_L^-)\}$$





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

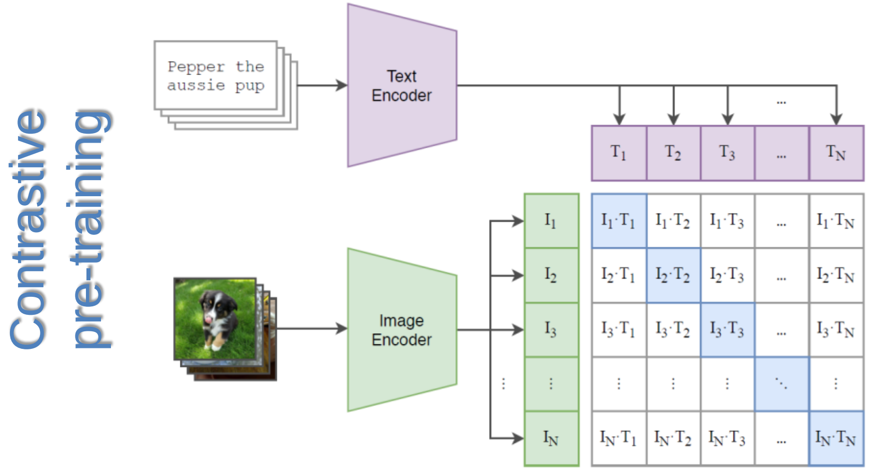
positive pairs

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

Similarity function can be cosine similarity

negative pairs and positive pairs

Positive and negative pairs:

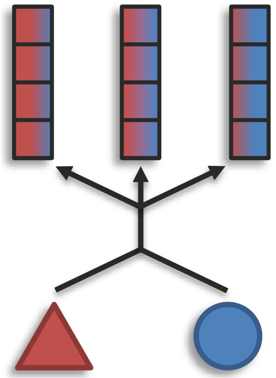


- CLIP encoders (f_L and f_V) are great for language-vision tasks
- $\boldsymbol{z}_L$ and $\boldsymbol{z}_V$ are coordinated but not identical representation spaces

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, arxiv 2021

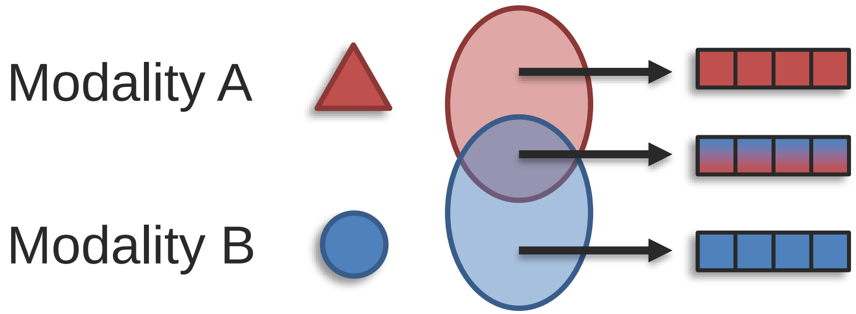
Representation Fission

Sub-Challenge 1c: Representation Fission

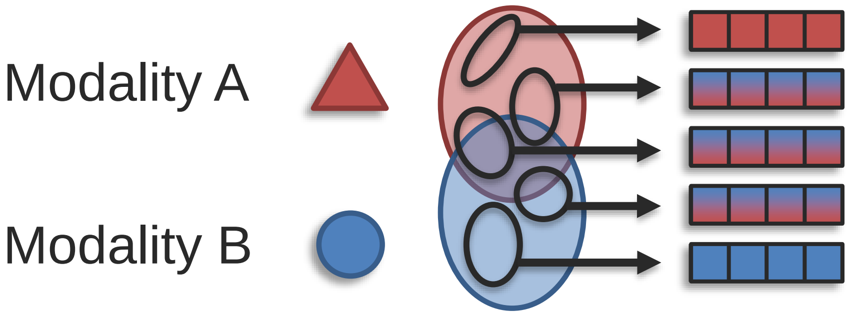


Definition: learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

Modality-level fission:



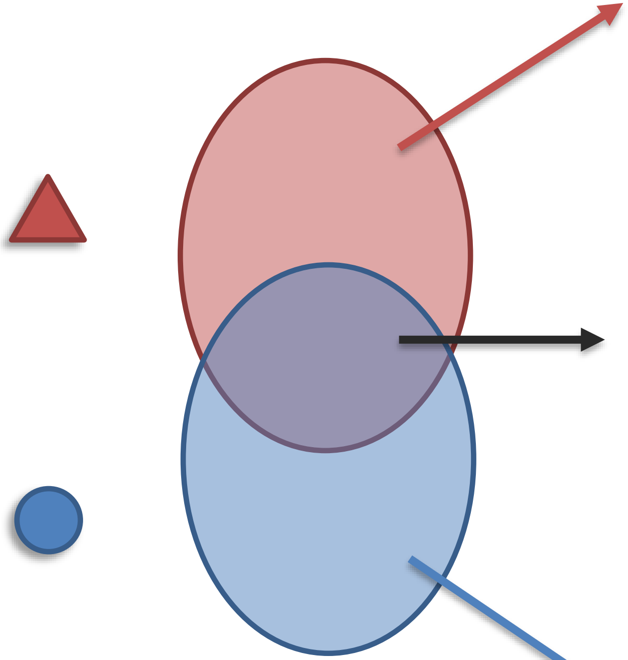
Fine-grained fission:



Modality-Level Fission

Language

Visual
(image)



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

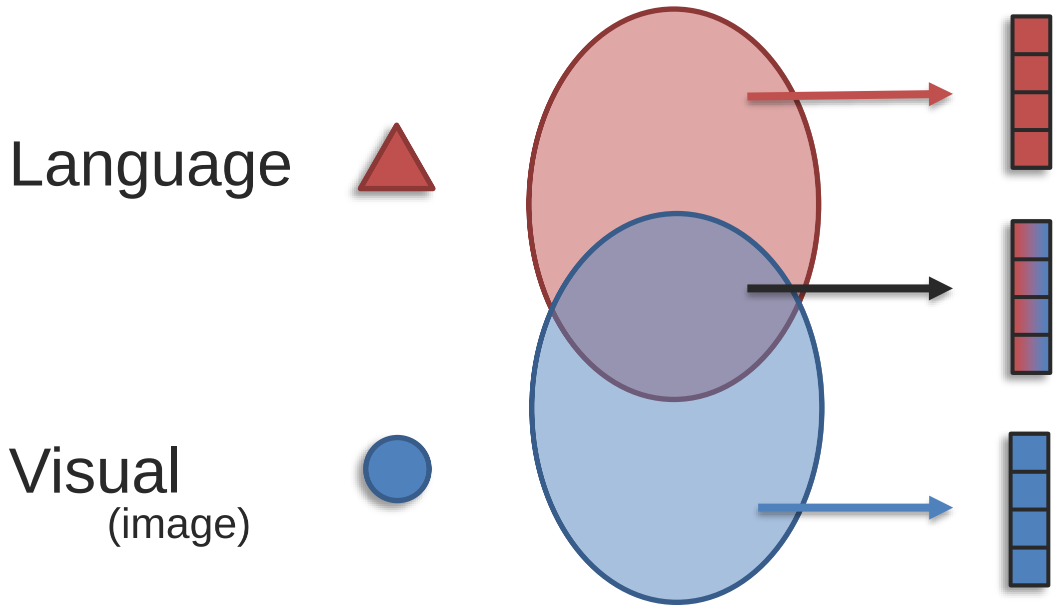
Information in both modalities

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

Information primarily in visual modality

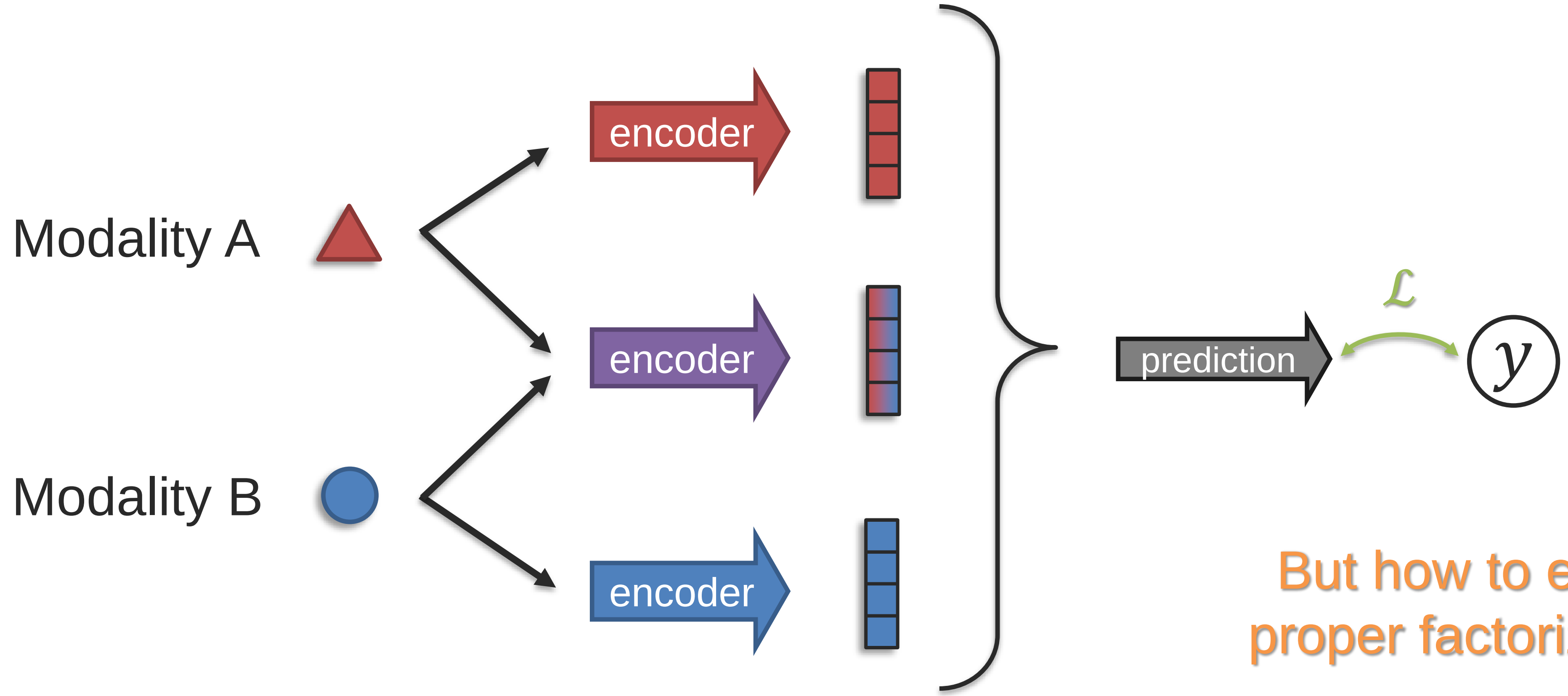
- Texture, visual appearance
- Depth, perspective, motion
- …

Modality-Level Fission



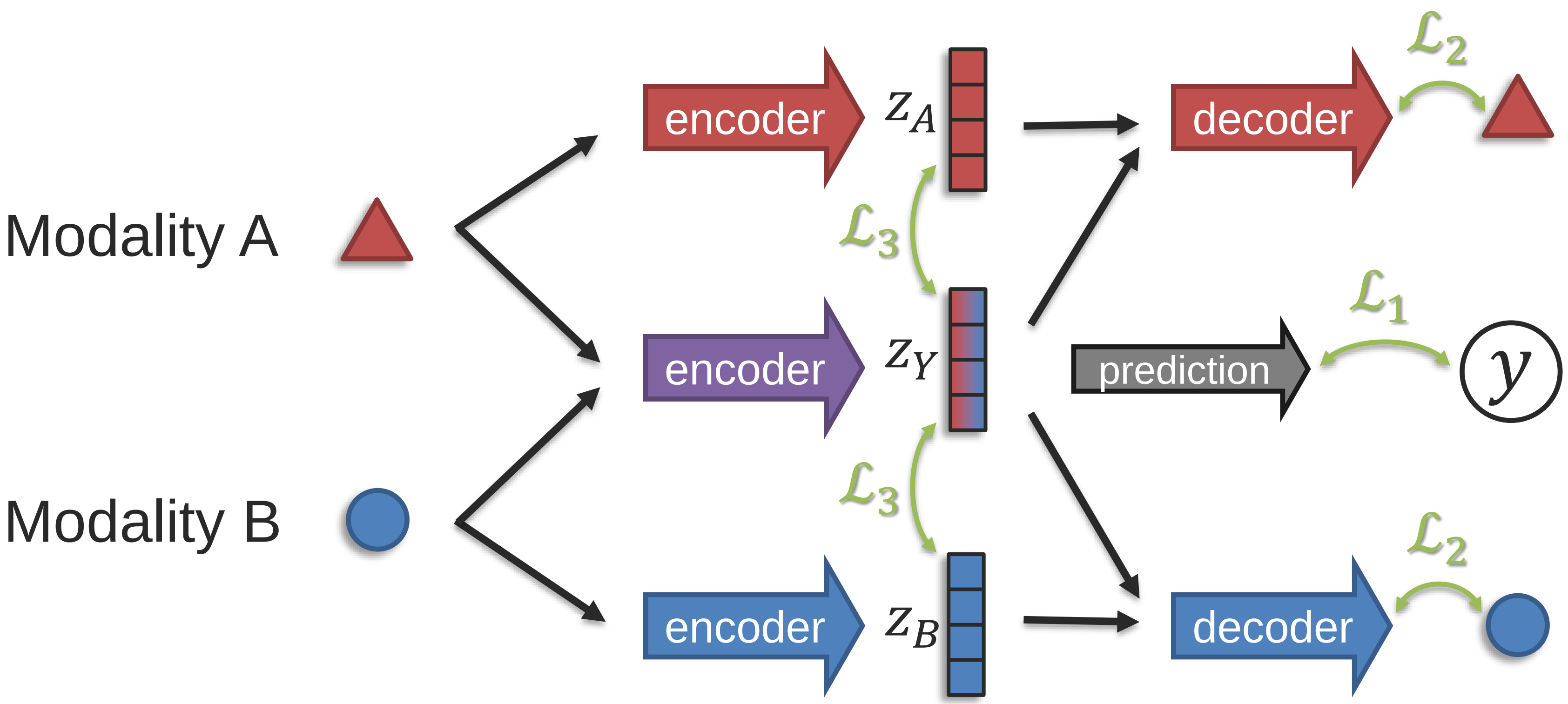
How to learn factorized multimodal representations?

A Discriminative Approach – Factorized Multimodal Representations



But how to ensure proper factorization?

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

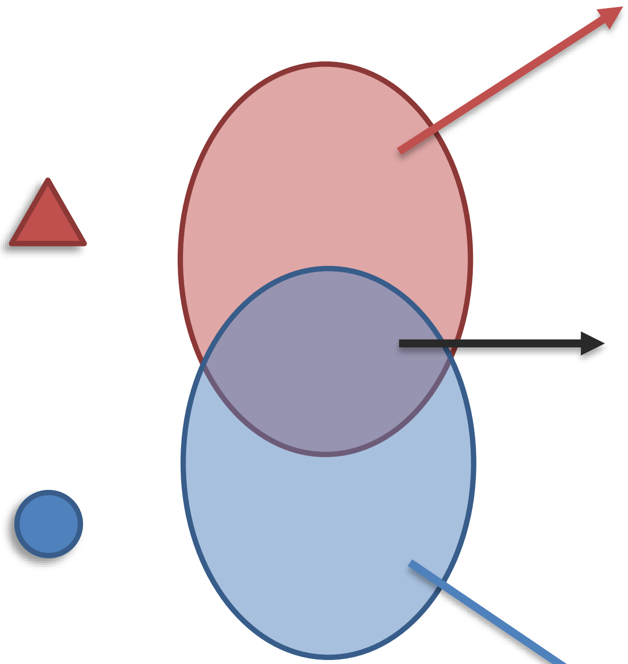
Separate priors for z_A, z_B and z_Y

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

Modality-Level Fission – Information Theory

Language

Visual
(image)



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

Information in both modalities

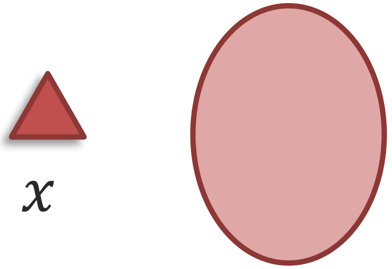
- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality

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

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 surprising its content is

x: "12, 34, 45, 62 was not a winning combination"

➡ Not surprising… So, low information

x: "11, 28, 38, 58 was a winning combination"

➡ Low chances… So, higher information

Information content $I(x)$

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

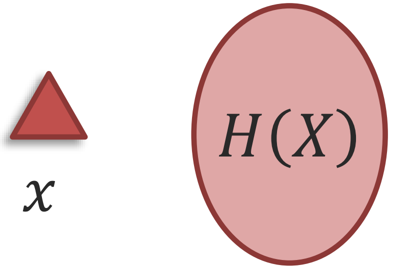
➡ But how to scale?

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

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

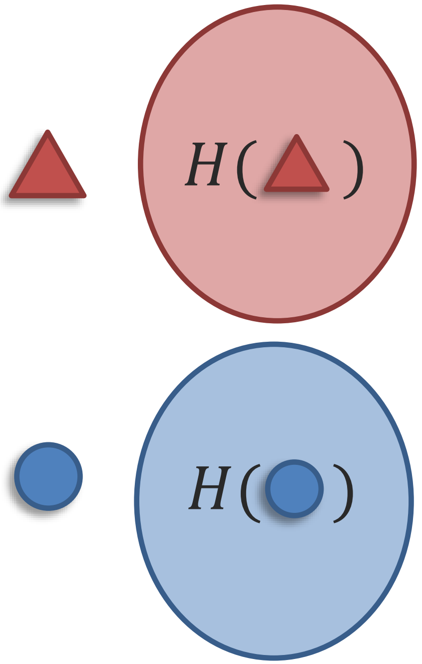
Entropy with Two Modalities

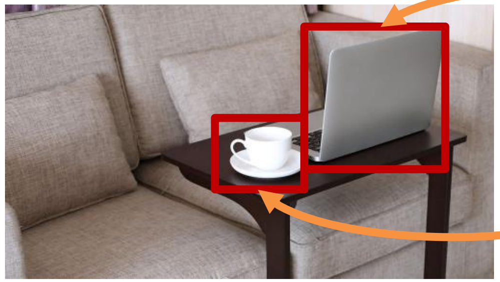
If no overlapping information

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

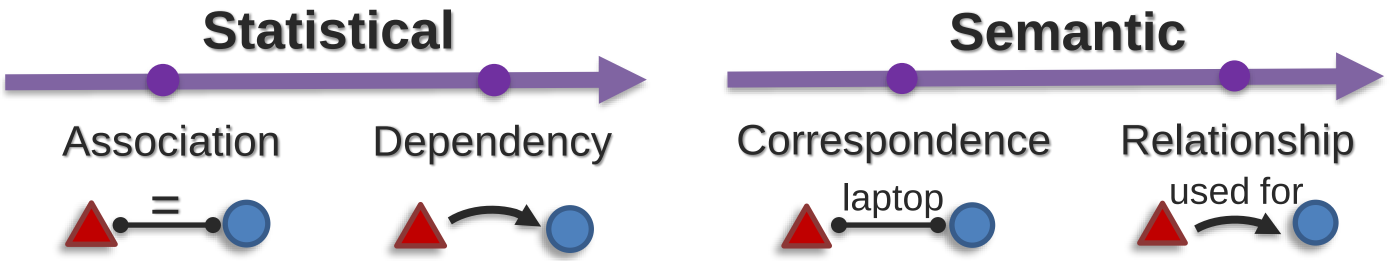
Modality A

Modality B





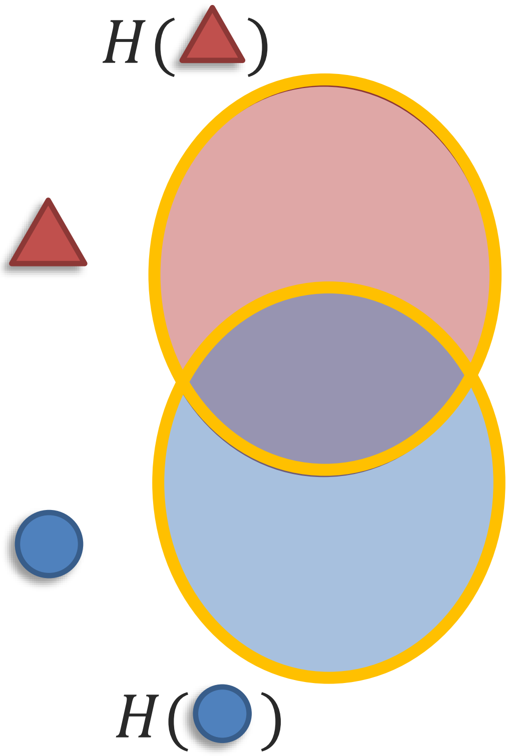
*A **teacup** on the right of a **laptop** in a clean room.*



Entropy with Two Modalities

Modality A

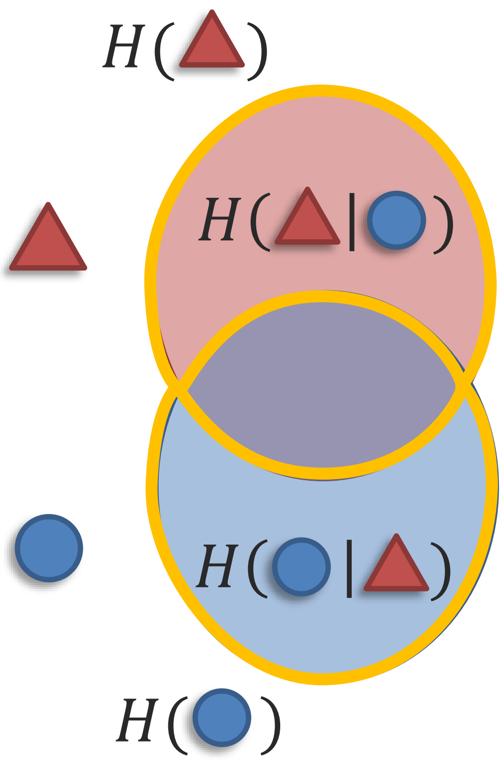
Modality B



Entropy with Two Modalities

Modality A

Modality B



Conditional entropy $H(Y|X)$

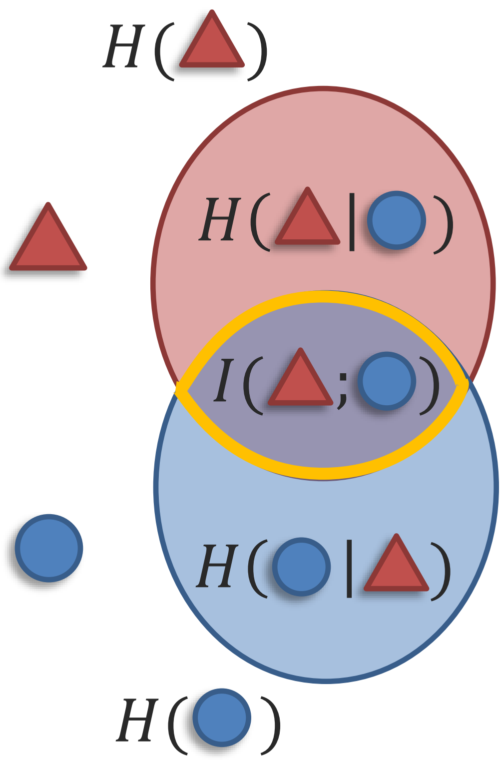
$$H(Y|X) = -\mathbb{E}_{X,Y}[\log p(y|x)]$$

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

Entropy with Two Modalities

Modality A

Modality B



Mutual information $I(X;Y)$

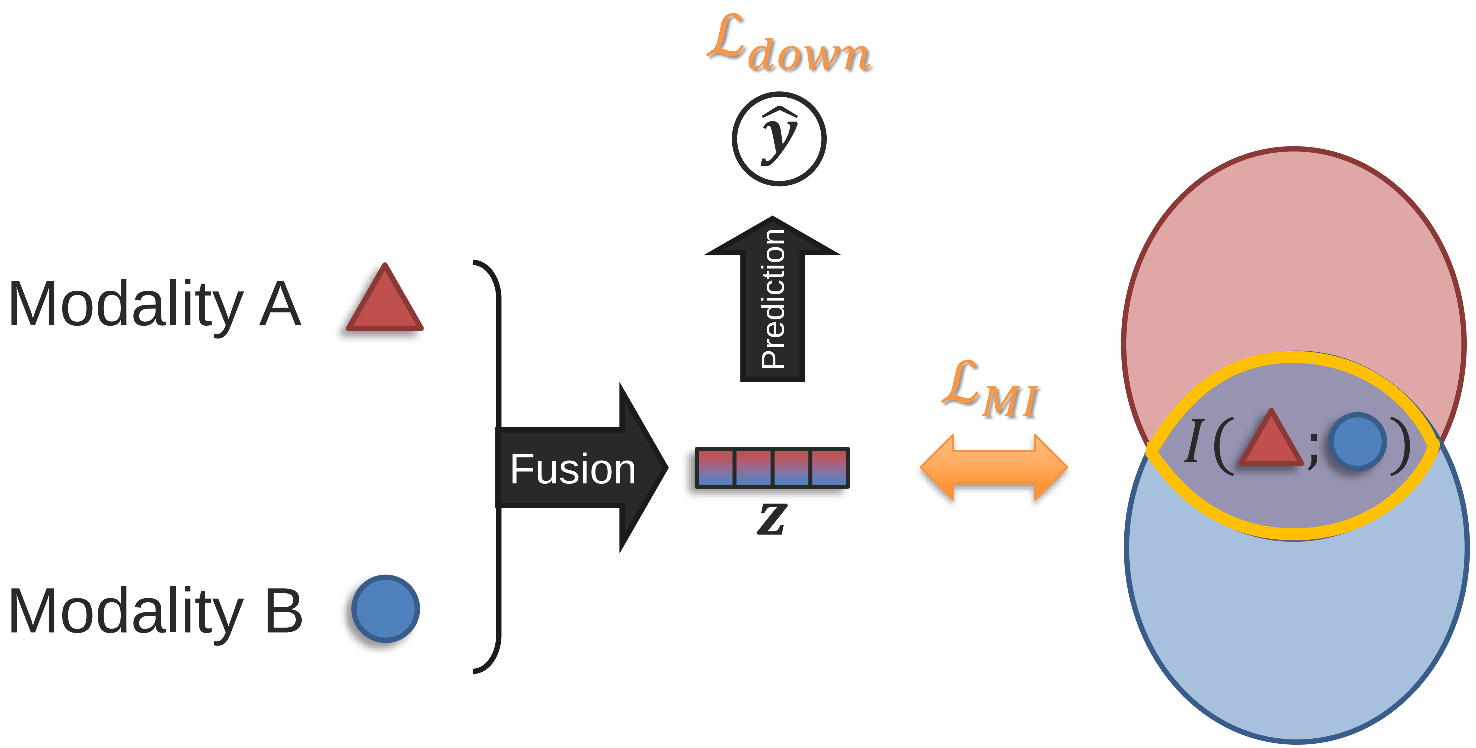
$$I(X;Y) = H(X) - H(X|Y)$$

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

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

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

Multimodal Fusion with Mutual Information

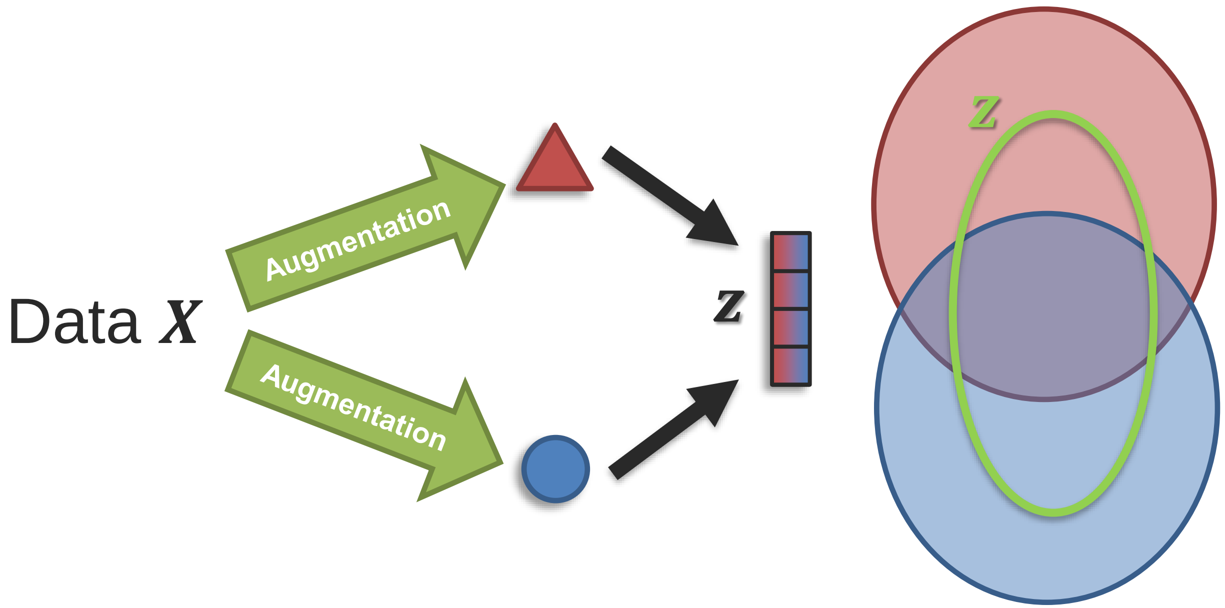


Assumption?

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

Colombo et al., Improving Multimodal fusion via Mutual Dependency Maximisation, EMNLP 2021

Link with Self-Supervised Learning



1. Maximize the mutual information

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

➡ Related to contrastive learning

2. Minimize the conditional entropy

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

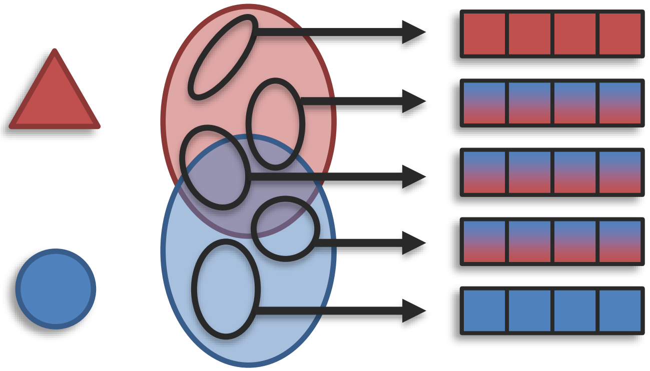
Information theory gives us a path towards disentangled representation learning

Tsai et al., Self-Supervised Learning from a Multi-View Perspective, ICLR 2021

Fine-Grained Fission

Modality A

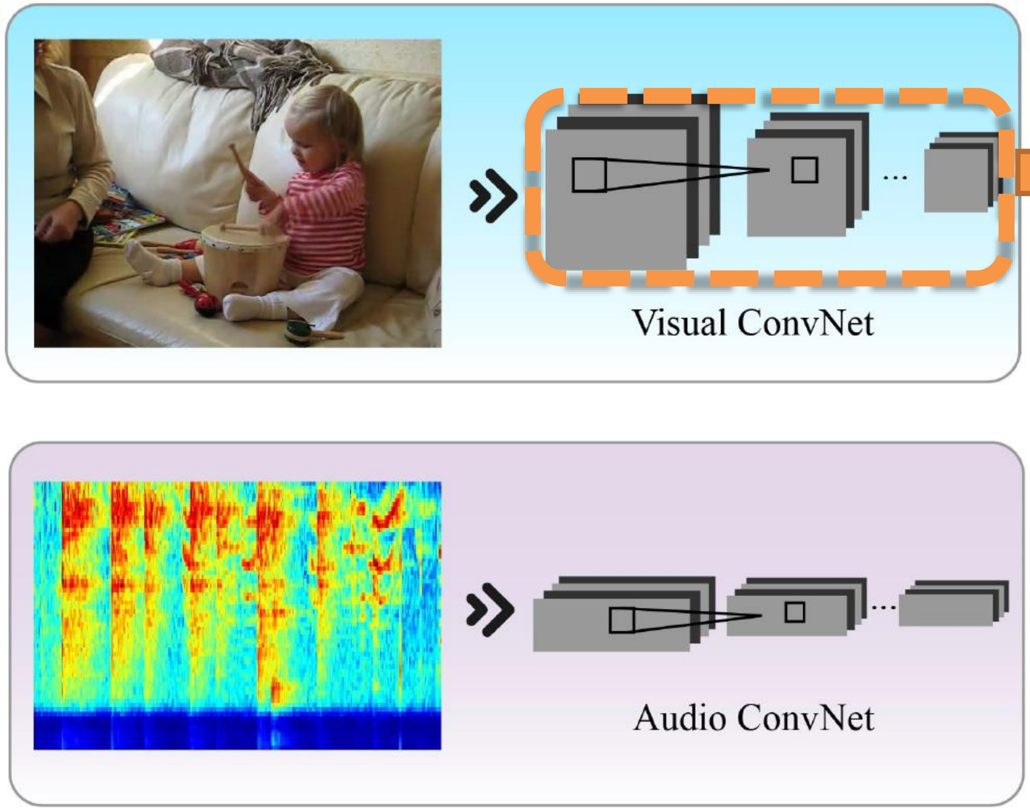
Modality B



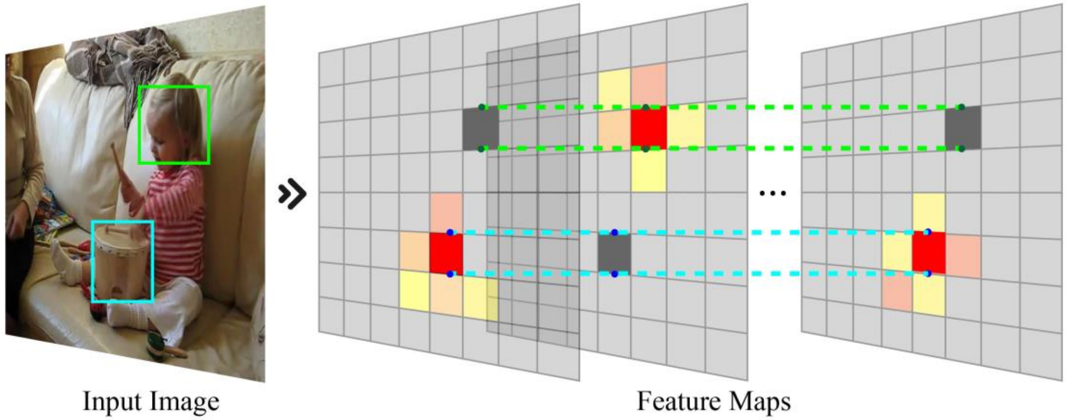
How to automatically discover these internal clusters, factors?

Fine-Grained Fission – A Clustering Approach

Unimodal Encoders



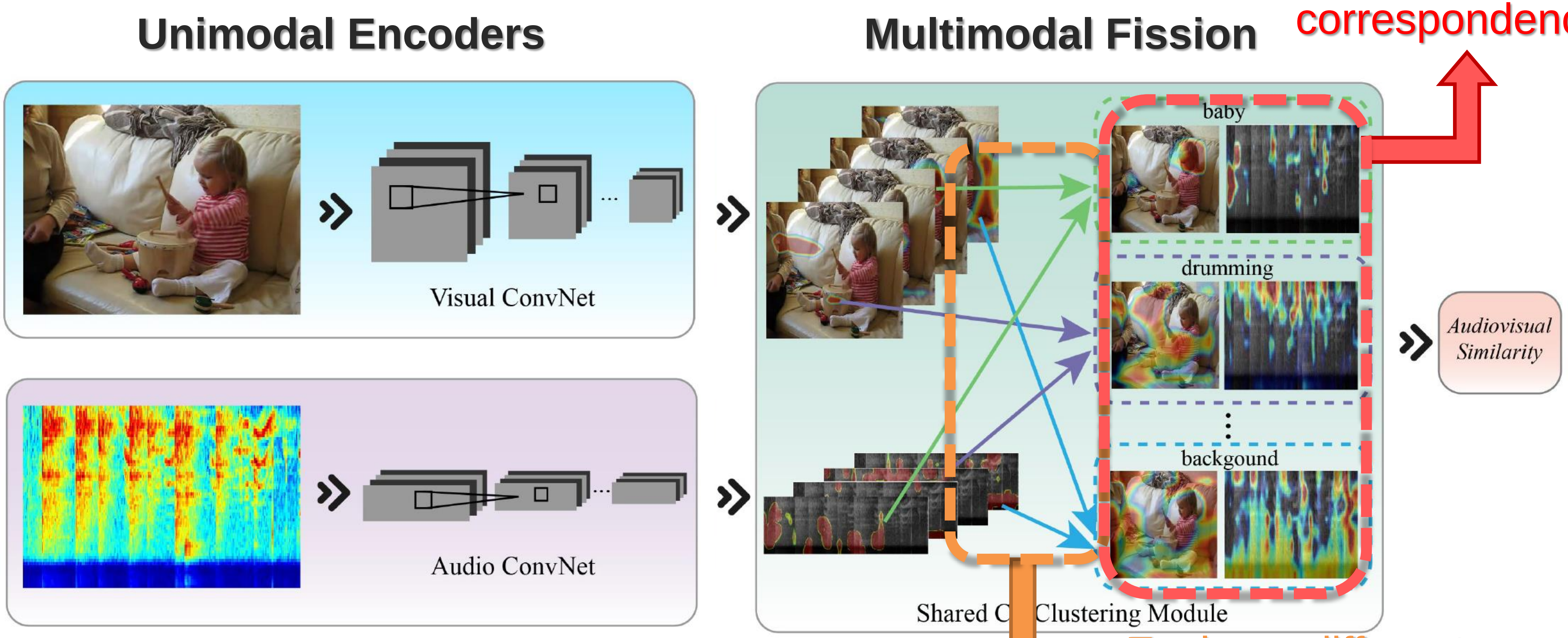
Localized activations for different objects



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)

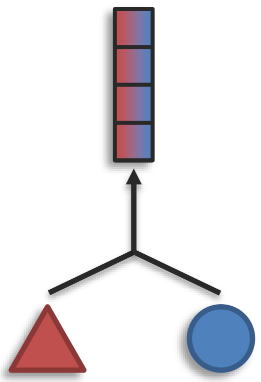
Hu et al., Deep Multimodal Clustering for Unsupervised Audiovisual Learning, CVPR 2019

Challenge 1: Representation

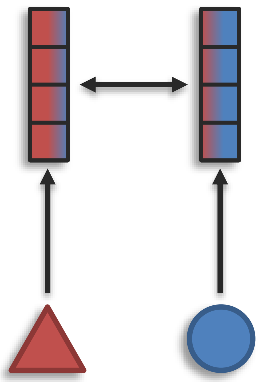
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

Fusion



Coordination



Fission