《多模态机器学习》

第五章 多模态表示

黄文炳
中国人民大学高瓴人工智能学院
hwenbing@126.com
2024年秋季

课程提纲

单模态表示

- 视觉模态
- 文本模态
- 三维点云
- 动作模态

基本概念

- 神经网络及其优化

经典多模态机器学习

- 多模态表示
- 多模态对齐
- 多模态推理
- 多模态生成
- 多模态迁移

通用多模态机器学习

- 通用多模态（大）模型
- 多模态预训练
- 多模态典型应用

内容提纲

① Cross-modal interactions

② Additive and multiplicative fusion

③ Gated fusion

内容提纲

① Cross-modal interactions

② Additive and multiplicative fusion

③ Gated fusion

Task 1: Representation (表示)

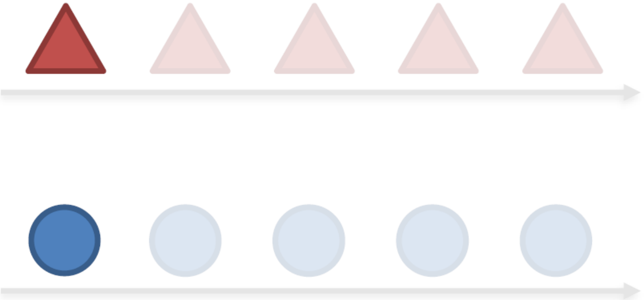
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

This is a core building block for most multimodal modeling problems!

Individual elements:

Modality A

Modality B



It can be seen as a "local" representation

or

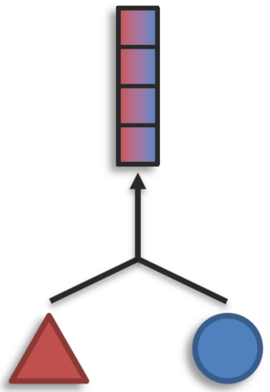
representation using holistic features

Task 1: Representation (表示)

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

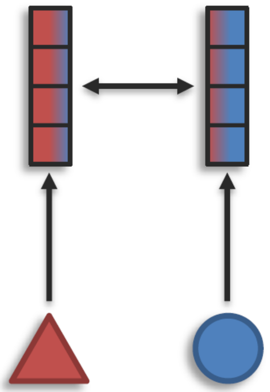
Sub-challenges:

Fusion



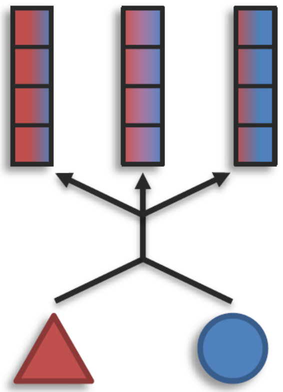
modalities **>** # representations

Coordination



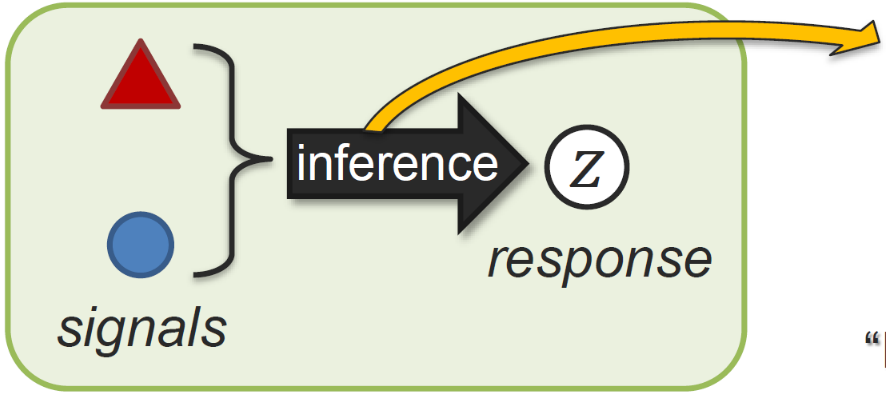
modalities **=** # representations

Fission



modalities **<** # representations

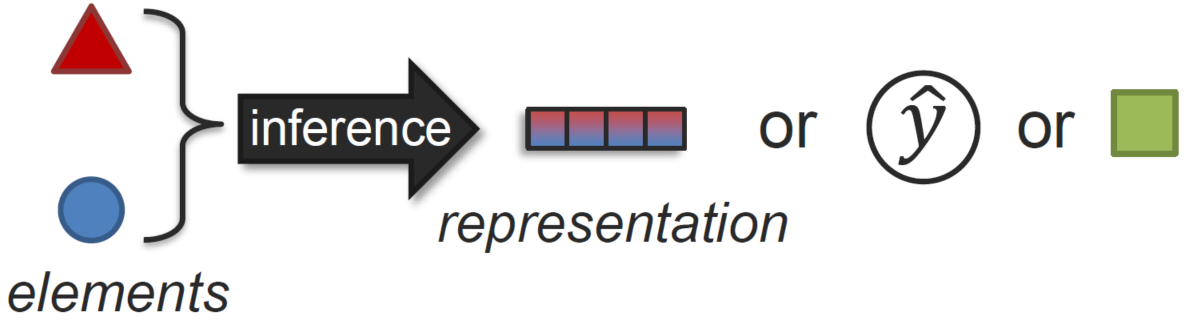
Cross-modal Interactions



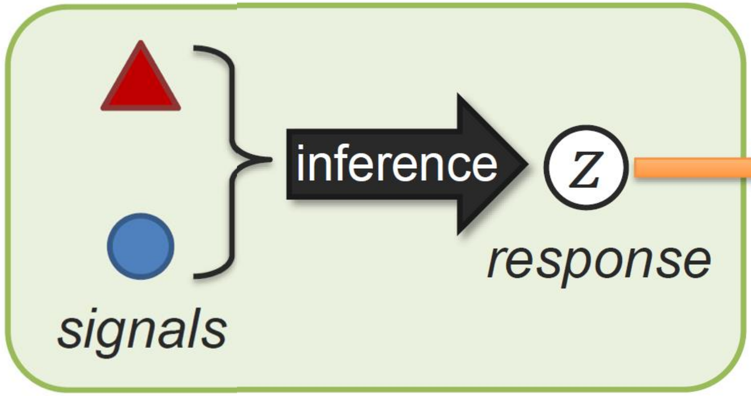
Interactions happen during inference!

“Inference” examples:

- Representation fusion
- Prediction task
- Modality translation

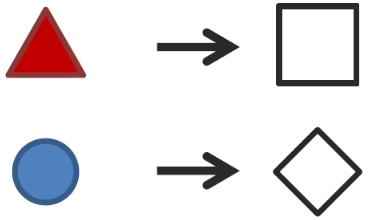


Cross-modal Interactions



Types of interaction responses?
(a taxonomy)

Unimodal
Non-redundancy



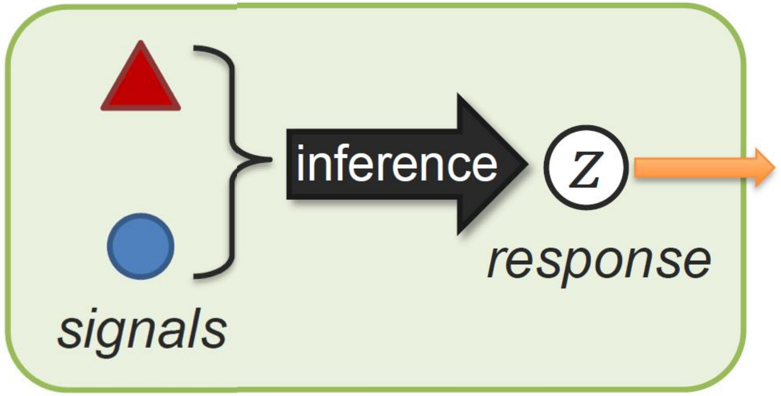
Is this a living room?

inference → *Yes!*

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

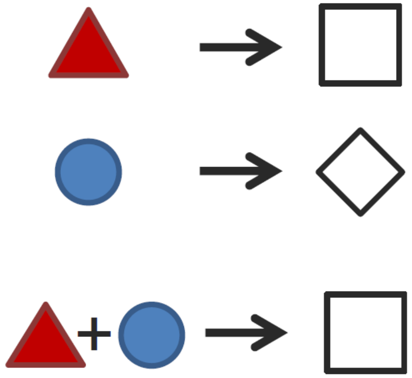
inference → *No, probably study room.*

Cross-modal Interactions



Types of interaction responses?
(a taxonomy)

Unimodal
Non-redundancy



Multimodal
dominance

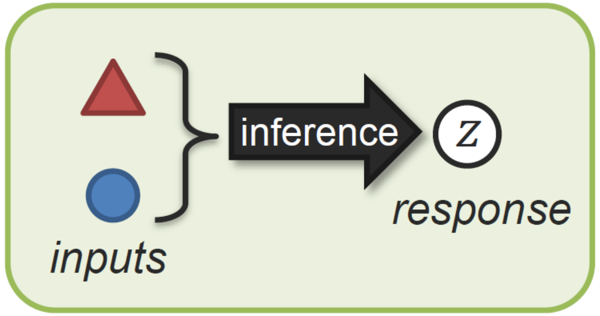
Is this a living room?

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

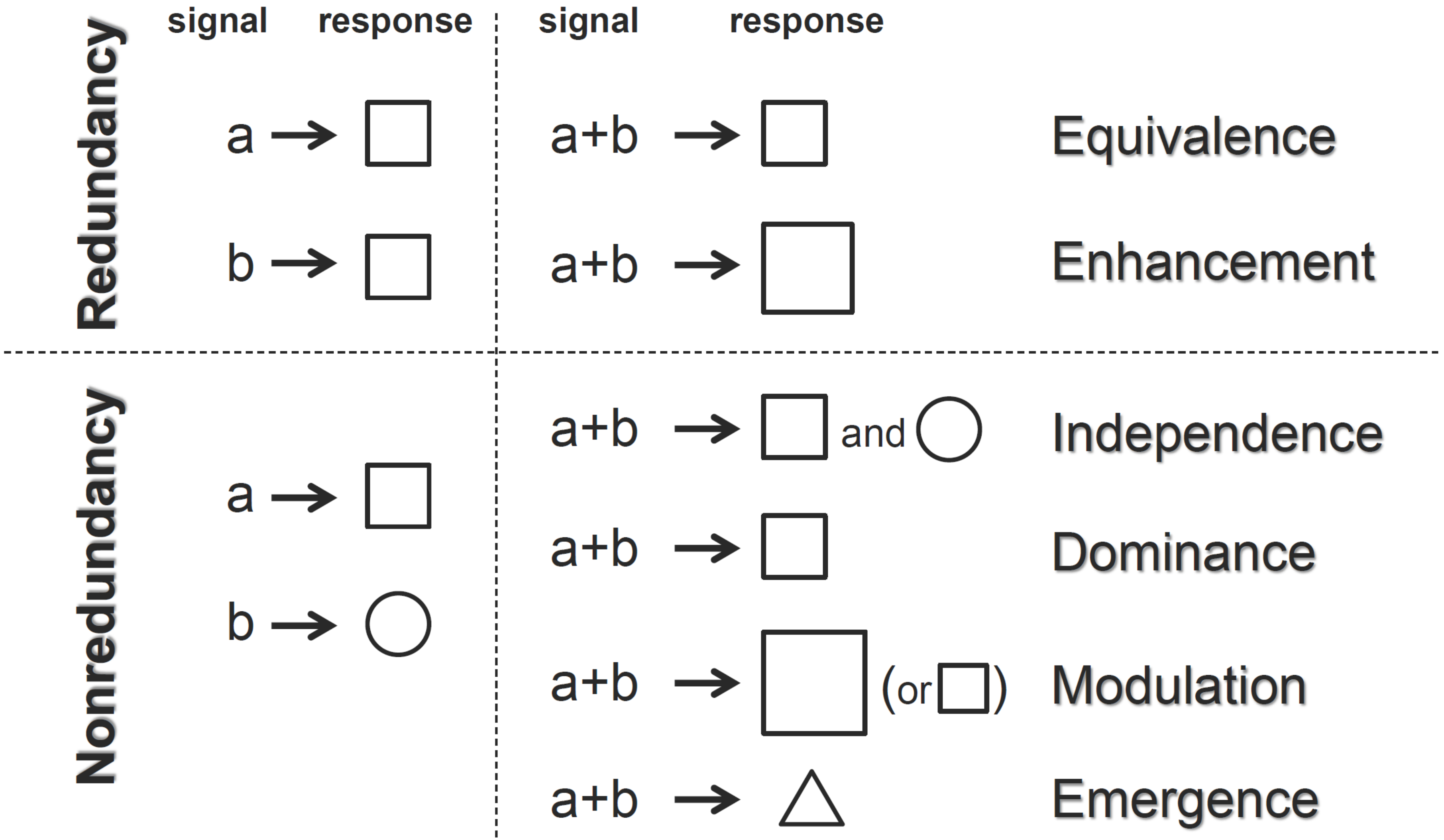
inference

Yes!

Taxonomy of Interaction Responses: A Behavioral Science View

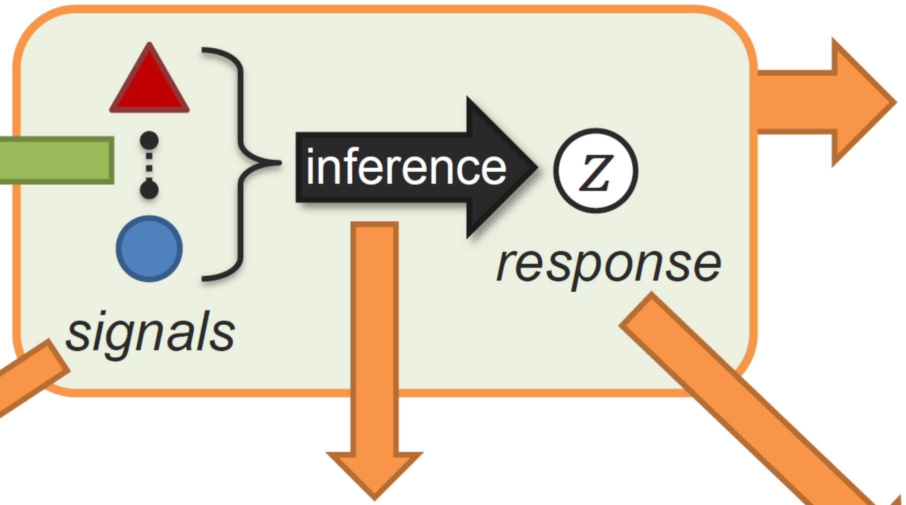


Multimodal Communication



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

Cross-modal Interactions



⑤ Connections

- Association
- Dependency
- Correspondence
- Relationship
- …

④ Context

- Structure context
- Task relevance
- Context dependence
- …

② Modalities

- Unimodal
- Bimodal
- Trimodal
- High-modal,
- …

③ Interactions

- Additive
- Multiplicative
- Polynomial
- Gated
- Nonlinear
- …

Today

① Responses

- Redundancy
- Non-redundancy
- Dominance
- Emergence
- …

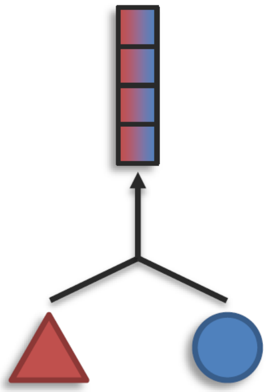
内容提纲

① Cross-modal interactions

② Additive and multiplicative fusion

③ Gated fusion

Sub-Challenge 1a: Representation Fusion



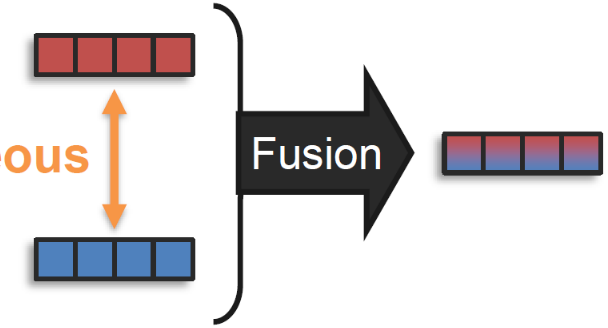
Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Basic fusion:

Modality A

Homogeneous

Modality B

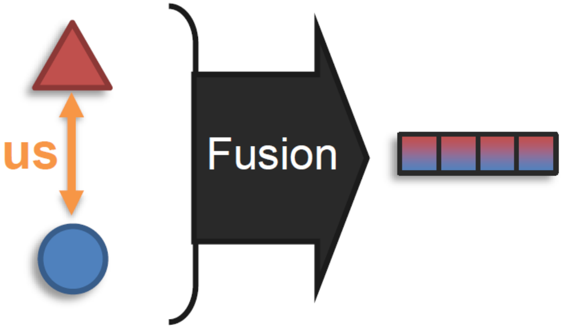


Raw-modality fusion:

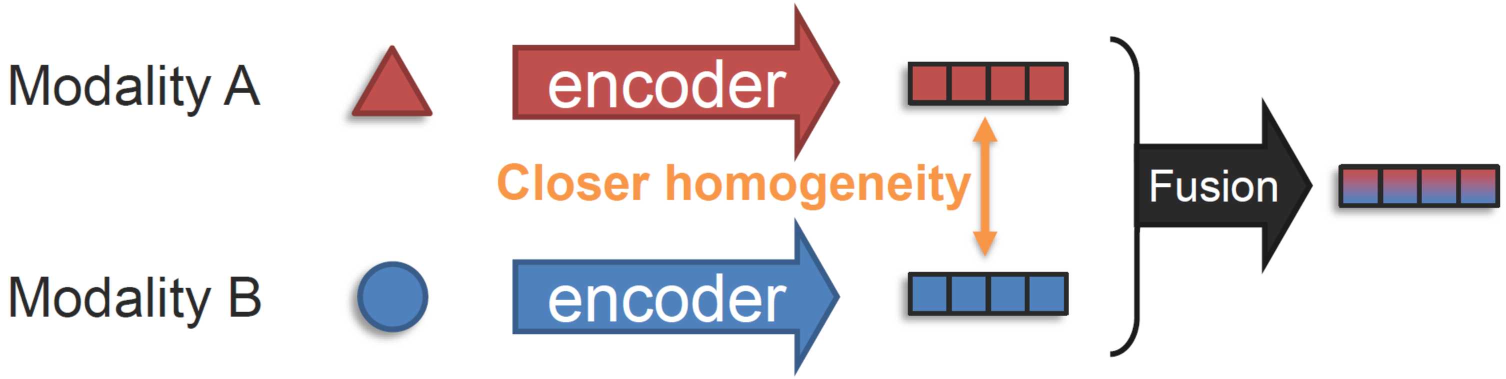
Modality A

Heterogeneous

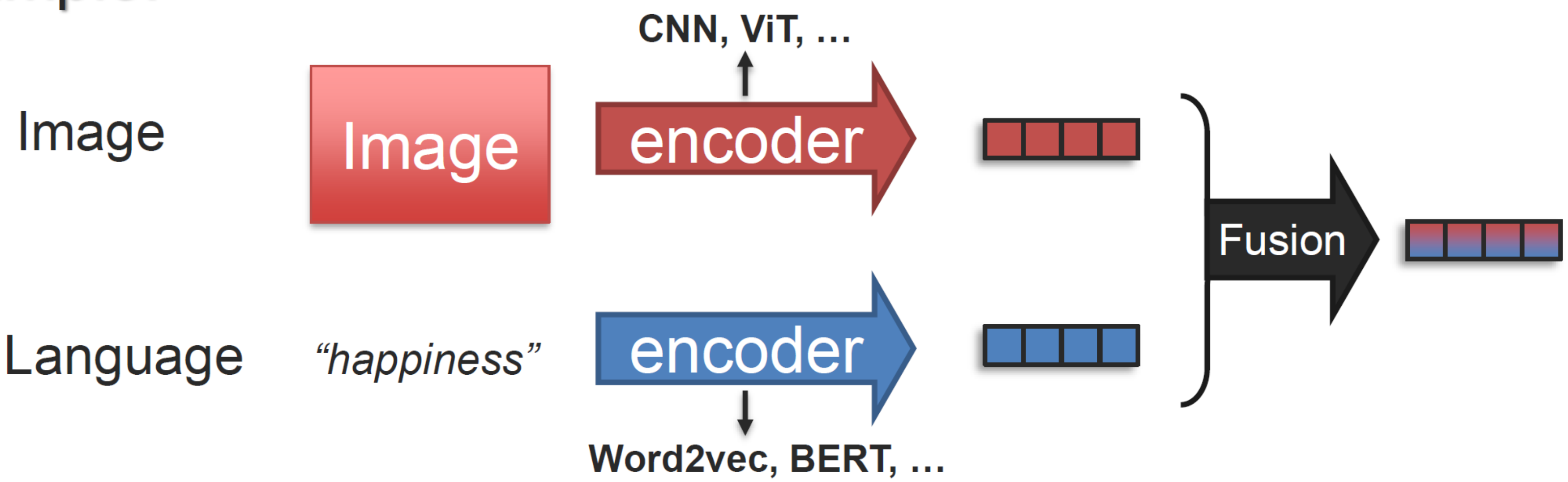
Modality B



Fusion with Unimodal Encoders

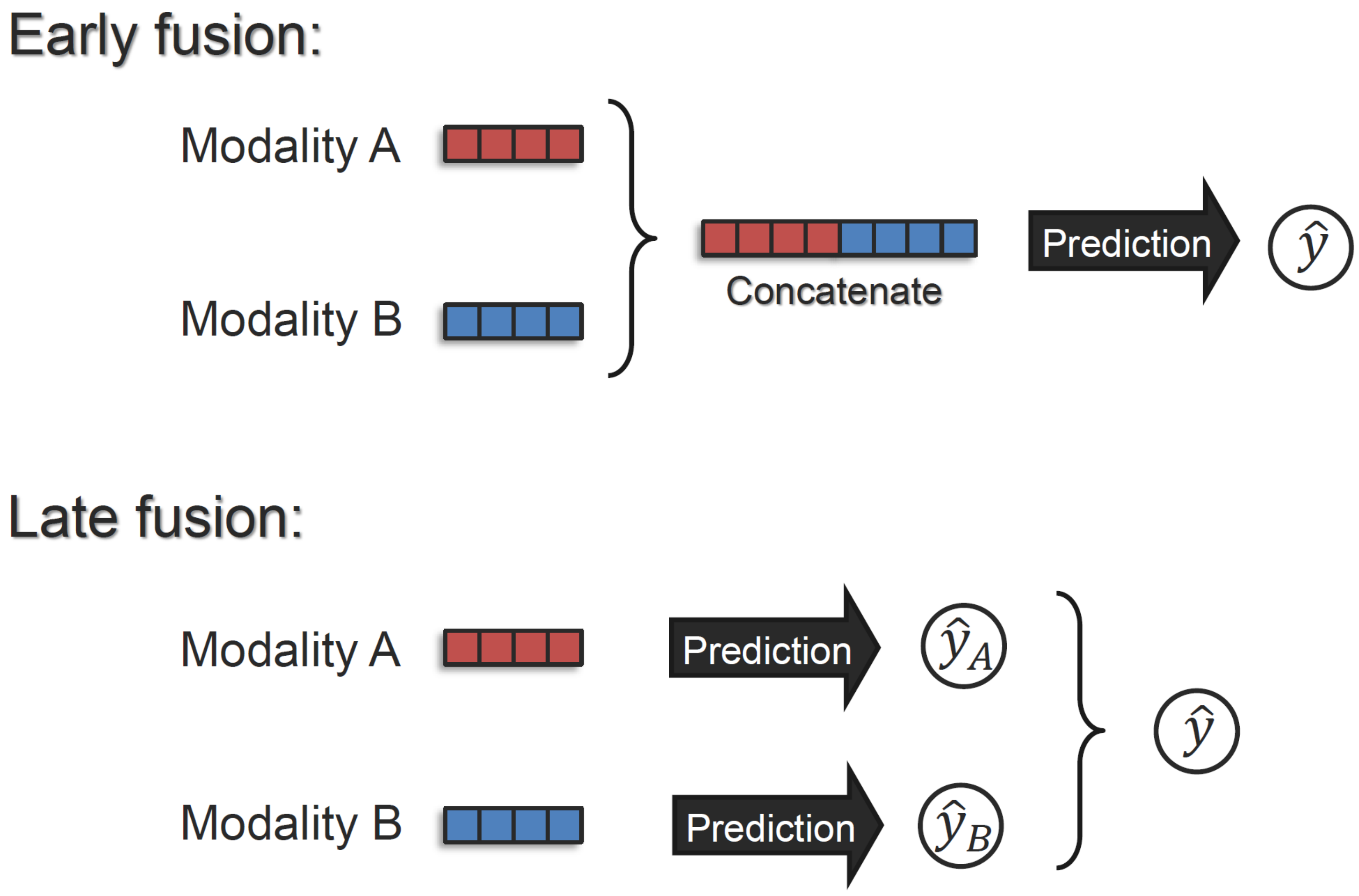


Example:

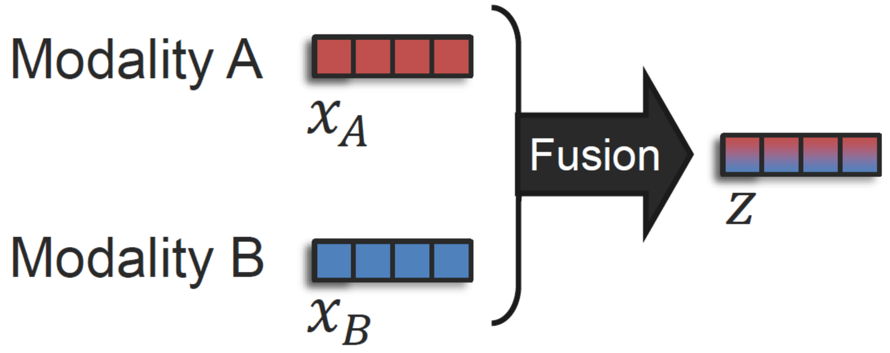


➡ Unimodal encoders can be jointly learned with fusion network, or pre-trained

Early and Late Fusion – A historical View



Basic Concepts for Representation Fusion (aka, Basic Fusion)



Goal: Model *cross-modal interactions* between the multimodal elements

➡ Let's study the univariate case first

↳ (only 1-dimensional features)

Linear regression:

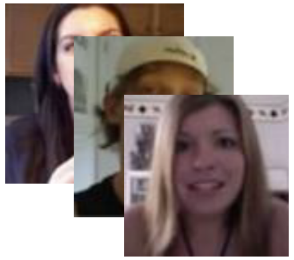
$$z = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

- w_0: intercept (bias term)
- $w_1 x_A + w_2 x_B$: Additive terms
- $w_3 (x_A \times x_b)$: Multiplicative term
- ϵ: error (residual term)

Linear Regression

Linear regression is used to test research hypotheses, over a whole dataset

300 book reviews



y: audience score

x_A: percentage of smiling

x_B: professional status
(0=non-critic, 1=critic)

***H1:** Does smiling reveal what the audience score was?*

***H2:** Does the effect of smiling depend on professional status?*

Linear regression:

$$y = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

- w_0: intercept (bias term)
- $w_1 x_A + w_2 x_B$: Additive terms
- $w_3 (x_A \times x_b)$: Multiplicative term
- ϵ: error (residual term)

w_0: average score when x_A and x_B are zero

w_1: effect from x_A variable only

w_2: effect from x_B variable only

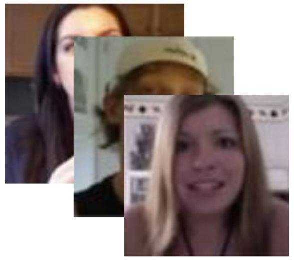
w_3: effect from x_A and x_B interaction only

ϵ: residual not modeled by w_0, w_1, w_2 or w_3

Linear Regression

Linear regression is used to test research hypotheses, over a whole dataset

300 book reviews



y: audience score

x_A: percentage of smiling

x_B: professional status (0=non-critic, 1=critic)

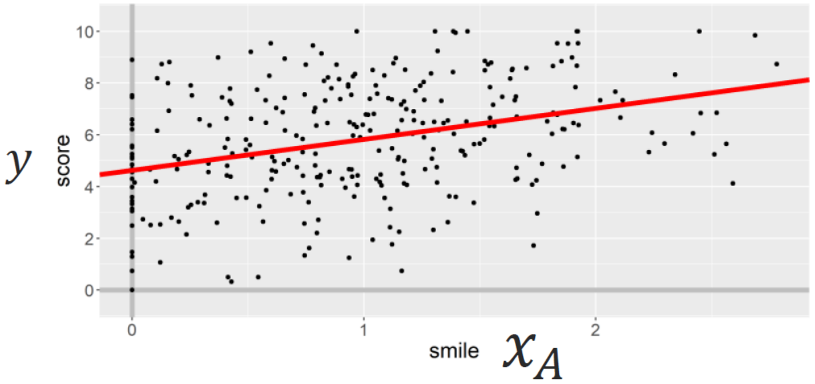
***H1:** Does smiling reveal what the audience score was?*

***H2:** Does the effect of smiling depend on professional status?*

Linear regression:

$$z = w_0 + w_1 x_A + \epsilon$$

slope



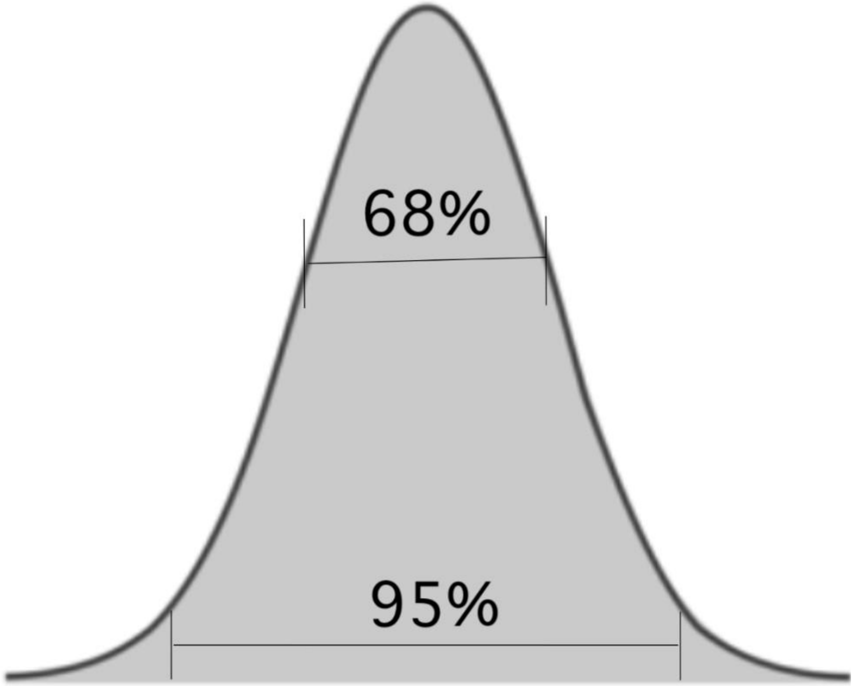
Confidence interval: *"95% confident that w parameter is contained within this interval"*

	Estimate	95% CI
w_0	4.63	[4.20, 5.06]
w_1	1.20	[0.83, 1.57]

Confidence interval does not contain 0, so effect is significant

p-values would be another way to test hypothesis

Confidence Interval



1. 已知总体标准差的情况

当总体标准差已知时，使用 **Z 分布** 来计算。

公式：

$$CI = \bar{X} \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

- $\bar{X}$: 样本均值
- $Z_{\alpha/2}$: Z 值，对应给定的置信水平（例如，95% 置信水平时，$Z_{\alpha/2} = 1.96$）
- σ: 总体标准差
- n: 样本大小

2. 未知总体标准差的情况

当总体标准差未知时，使用 **t 分布** 来计算。

公式：

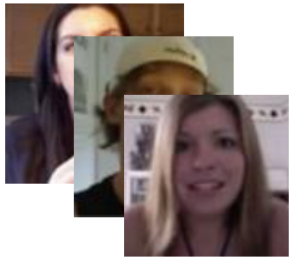
$$CI = \bar{X} \pm t_{\alpha/2,df} \cdot \frac{s}{\sqrt{n}}$$

- $\bar{X}$: 样本均值
- $t_{\alpha/2,df}$: t 值，基于样本大小 n 的自由度 $df = n - 1$ 查表

Linear Regression

Linear regression is used to test research hypotheses, over a whole dataset

300 book reviews



y: audience score

x_A: percentage of smiling

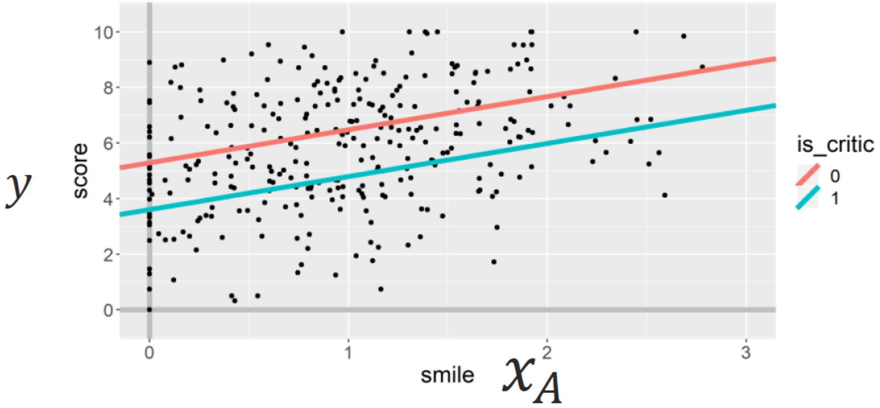
x_B: professional status (0=non-critic, 1=critic)

H1: Does smiling reveal what the audience score was?

H2: Does the effect of smiling depend on professional status?

Linear regression:

$$z = w_0 + \boxed{w_1} x_A + \boxed{w_2} x_B + \epsilon$$



	Estimate	95% CI
w_0	5.29	[4.86, 5.73]
w_1	1.19	[0.85, 1.53]
w_2	−1.69	[−2.14, −1.24]

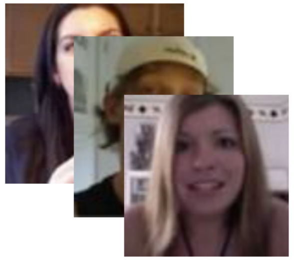
Positive effect (w_1)

Negative effect (w_2)

Linear Regression

Linear regression is used to test research hypotheses, over a whole dataset

300 book reviews



y: audience score

x_A: percentage of smiling

x_B: professional status

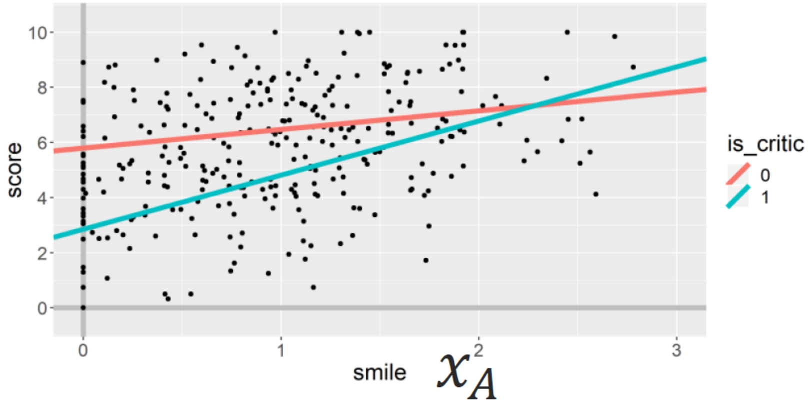
(0=non-critic, 1=critic)

***H1:** Does smiling reveal what the audience score was?*

***H2:** Does the effect of smiling depend on professional status?*

Linear regression:

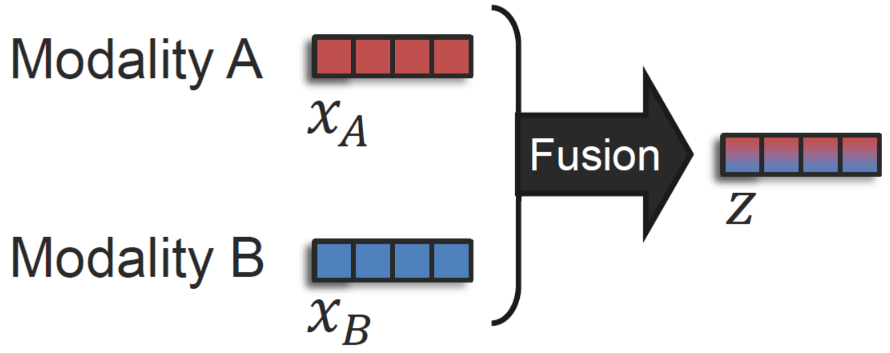
$$z = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$



	Estimate	95% CI
w_0	5.79	[5.29, 6.29]
w_1	0.68	[0.25, 1.11]
w_2	−2.94	[−3.73, −2.15]
w_3	1.29	[0.61, 1.97]

Multiplicative interaction!

Basic Concepts for Representation Fusion (aka, Basic Fusion)



Goal: Model *cross-modal interactions* between the multimodal elements

➡ Let's study the univariate case first

↳ (only 1-dimensional features)

Linear regression:

$$z = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

- w_0: intercept (bias term)
- $w_1 x_A + w_2 x_B$: Additive terms
- $w_3 (x_A \times x_b)$: Multiplicative term
- ϵ: error (residual term)

① Additive terms:

$$z = w_1 x_A + w_2 x_B + \epsilon$$

② Multiplicative "interaction" term:

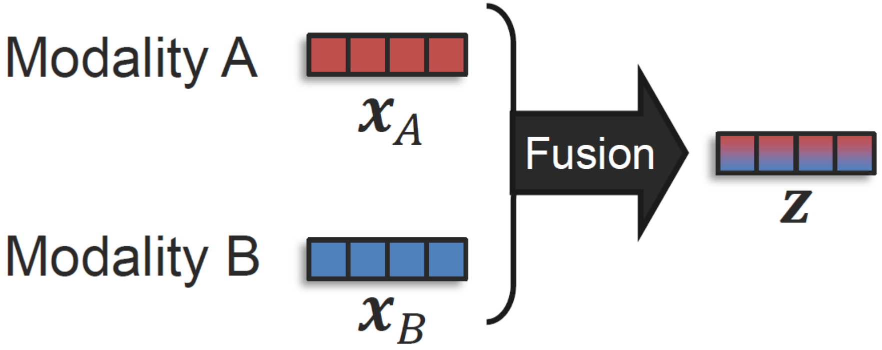
$$z = w_3 (x_A \times x_b) + \epsilon$$

③ Additive and multiplicative terms:

$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

Additive Fusion

Back to multivariate case!

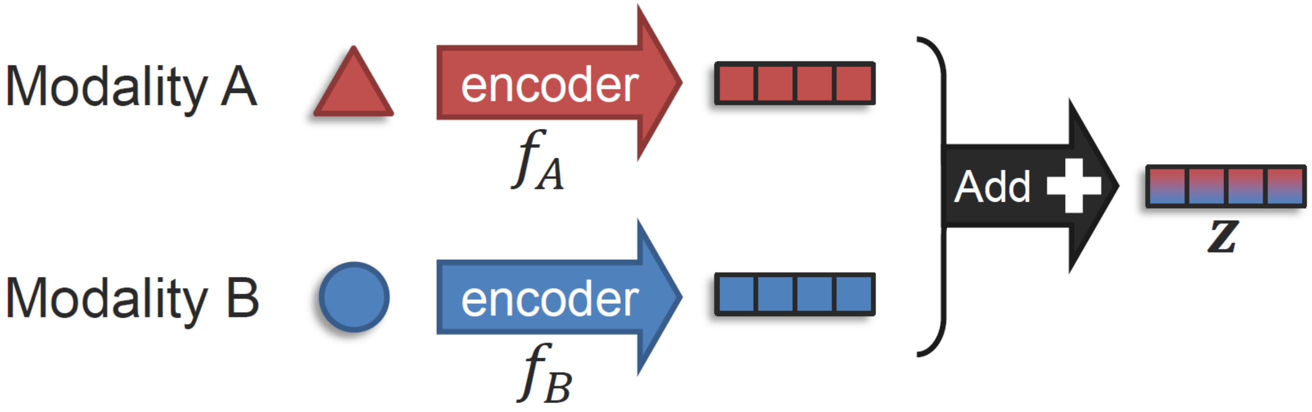


Additive fusion:

$$\boldsymbol{z} = \boldsymbol{W}_1 \boldsymbol{x}_A + \boldsymbol{W}_2 \boldsymbol{x}_B$$

1-layer neural network can be seen as additive

With unimodal encoders:

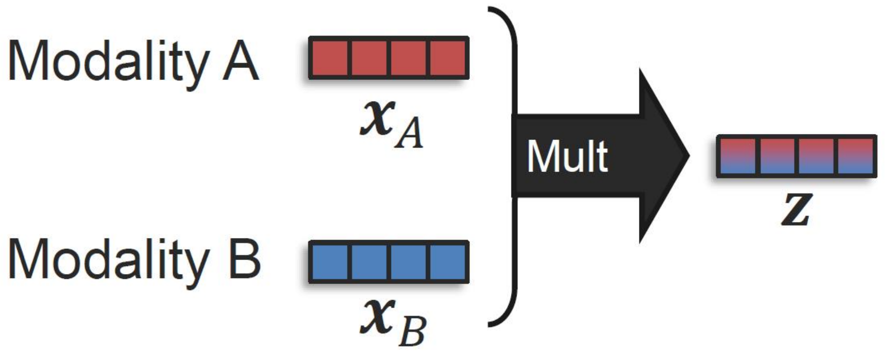


Additive fusion:

$$\boldsymbol{z} = f_A(\blacktriangle) + f_B(\bullet)$$

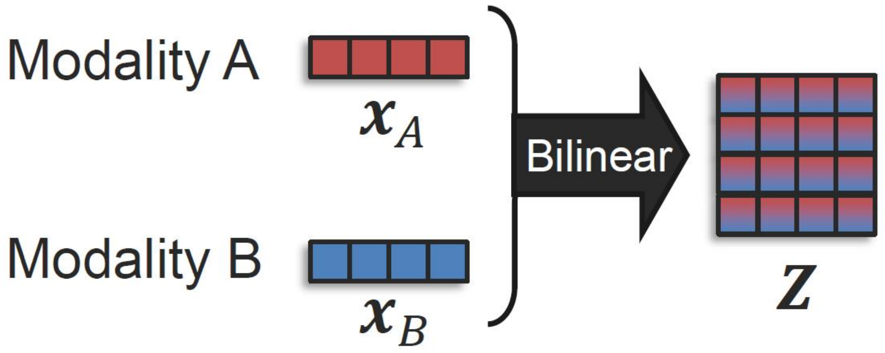
It could be seen as an ensemble approach (late fusion)

Multiplicative Fusion



Simple multiplicative fusion:

$$z = x_A \odot x_B$$



Bilinear Fusion:

$$Z = x_A^\top x_B$$

$$\text{vec}(Z) = x_A \otimes x_B$$

Jayakumar et al., Multiplicative Interactions and Where to Find Them, ICLR 2020

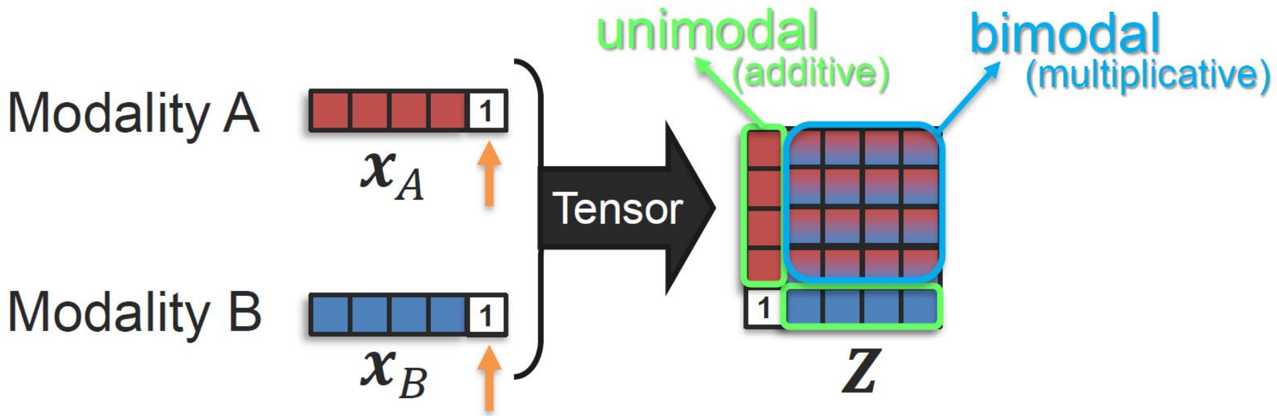
Kronecker product

If $\boldsymbol{A} \in \mathbb{R}^{m \times n}, \boldsymbol{B} \in \mathbb{R}^{p \times q}$, then the Kronecker product $\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{R}^{pm \times qn}$:

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix}$$

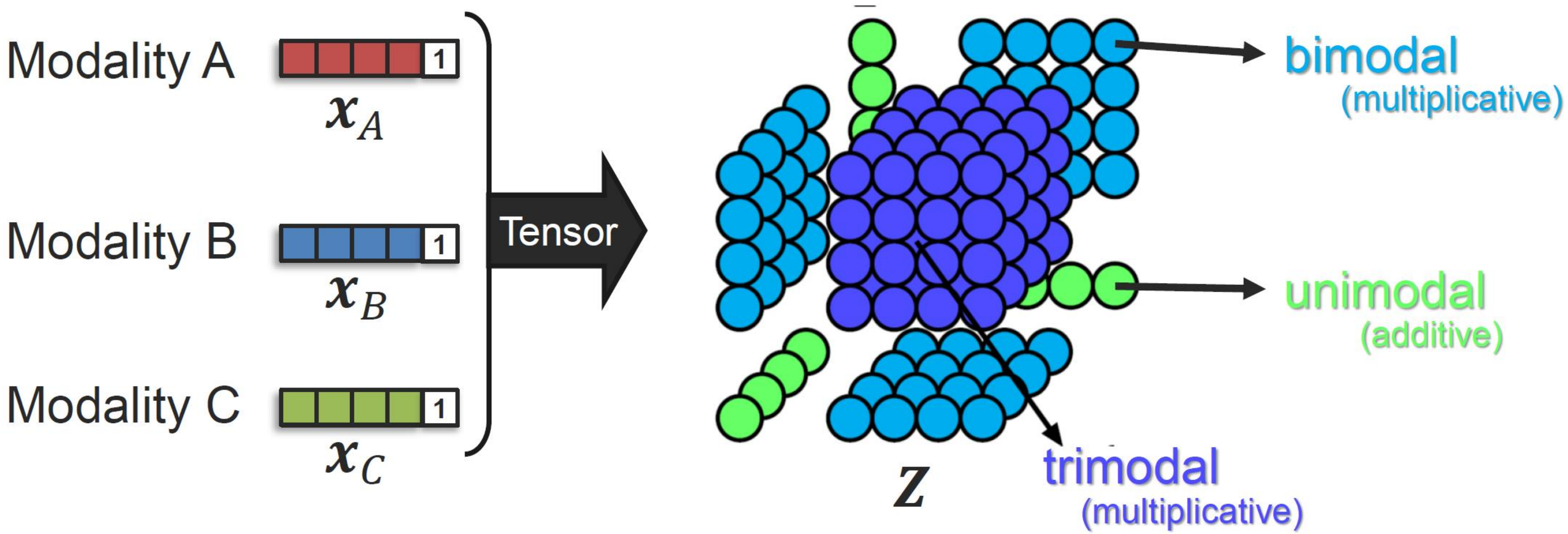
Jayakumar et al., Multiplicative Interactions and Where to Find Them, ICLR 2020

Multiplicative Fusion



Tensor Fusion (bimodal):

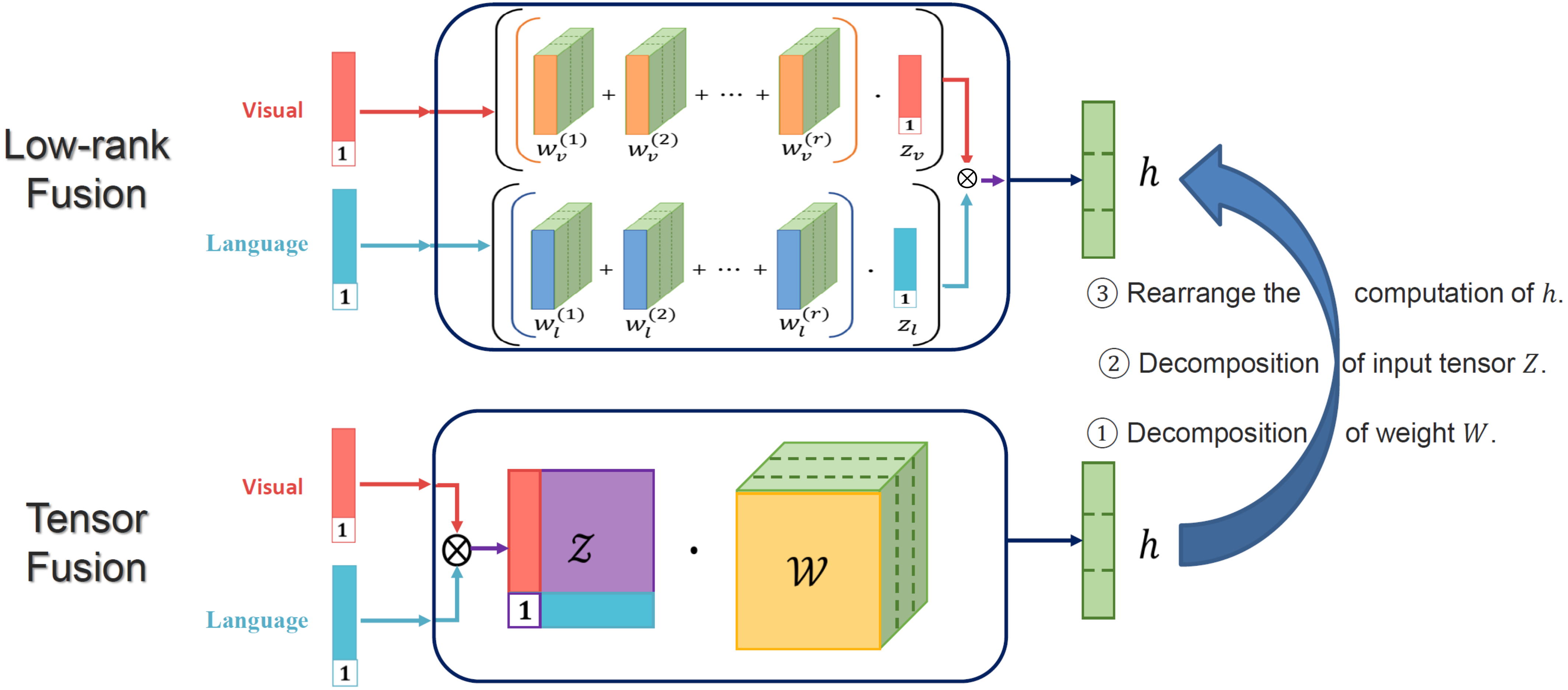
$$Z = [x_A \ 1]^\top [\ x_B \ 1]$$



… but the weight matrix may end up quite large!

Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis, EMNLP 2017

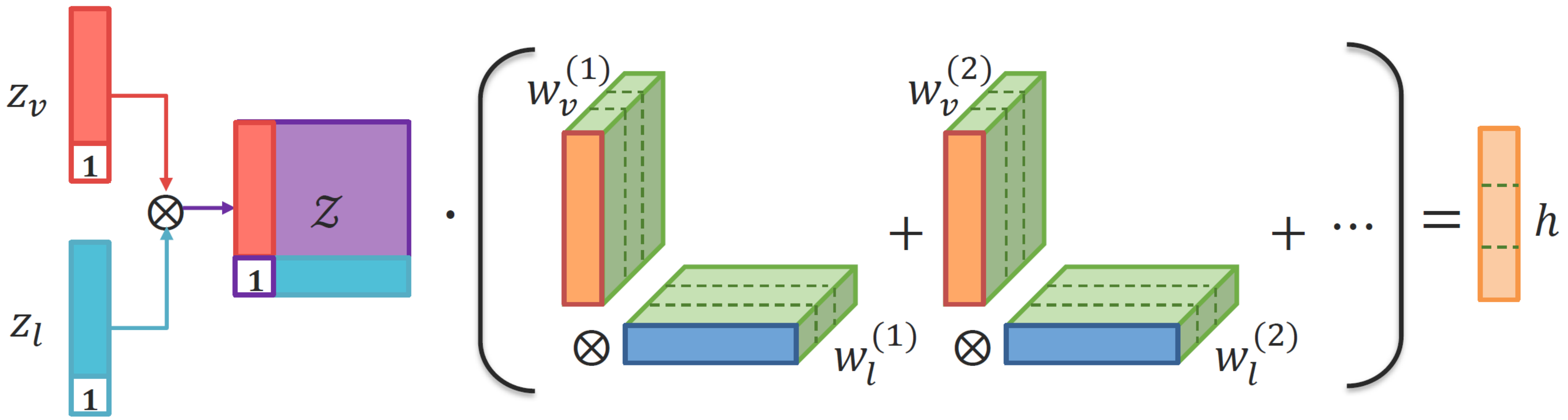
Low-rank Fusion



Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

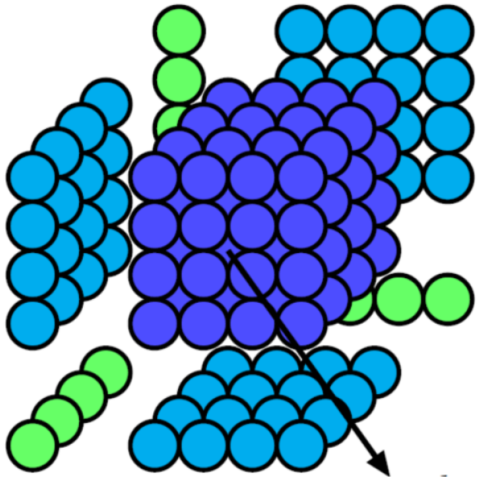
Low-rank Fusion

$$\mathrm{vec}(\boldsymbol{Z}) \cdot \mathrm{vec}(\boldsymbol{W})$$
$$= (\boldsymbol{z}_v \otimes \boldsymbol{z}_l) \cdot (\sum_{k=1}^{r} \boldsymbol{w}_v^{(k)} \otimes \boldsymbol{w}_l^{(k)})$$
$$= \sum_{k=1}^{r} \boldsymbol{z}_v \cdot \boldsymbol{w}_v^{(k)} \otimes \boldsymbol{z}_l \cdot \boldsymbol{w}_l^{(k)}$$

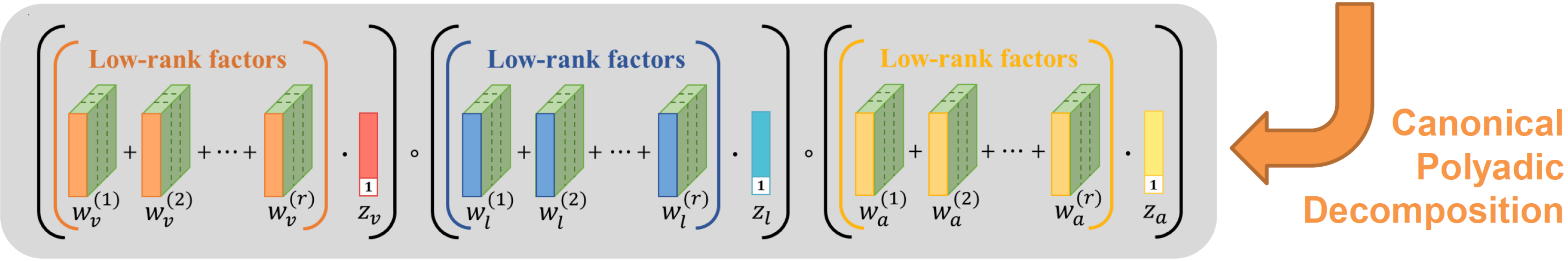


Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

Low-rank Fusion with Trimodal Input



Low-rank Fusion :



Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

Going Beyond Additive and Multiplicative Fusion

Additive interaction:

$$z = w_1 x_A + w_2 x_B$$

⟵ First-order polynomial

Additive and multiplicative interaction:

$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_B)$$

⟵ Second-order polynomial

Trimodal fusion (e.g., tensor fusion):

$$z = \underbrace{w_1 x_A + w_2 x_B + w_3 x_C}_{\text{Unimodal terms (first-order)}} + \underbrace{w_4 (x_A \times x_C) + w_5 (x_A \times x_C) + w_6 (x_B \times x_C)}_{\text{Bimodal terms (second-order)}} + \underbrace{w_7 (x_A \times x_B \times x_C)}_{\text{Trimodal terms (third-order)}}$$

Unimodal terms (first-order)

Bimodal terms (second-order)

Trimodal terms (third-order)

Can we add higher-order interaction terms?

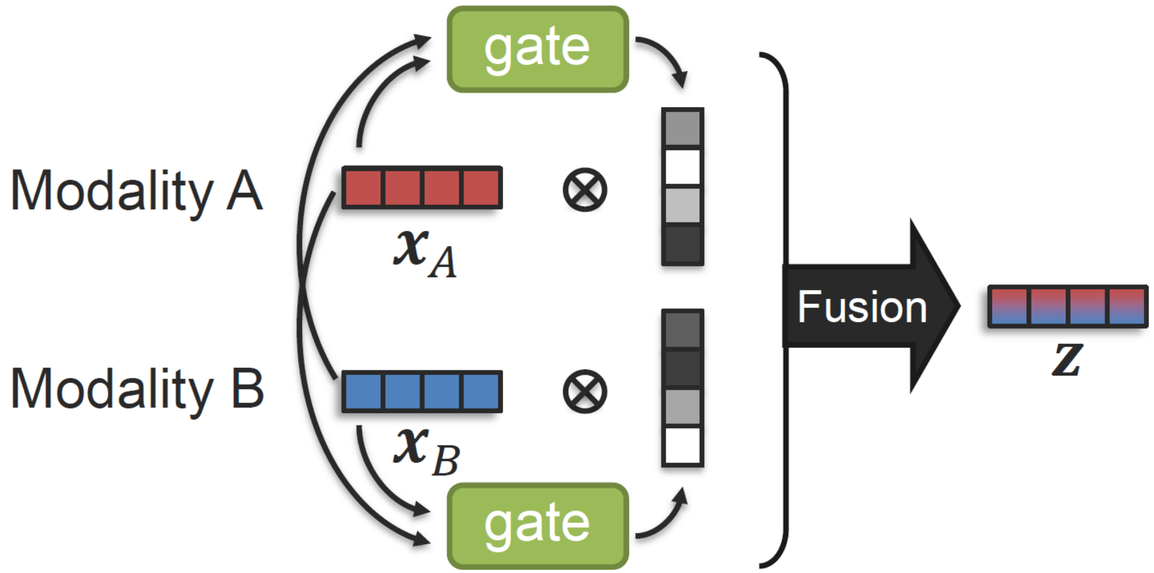
For example:

$$+w_8 (x_A^2 \times x_B^2 \times x_C^2)$$

$$+w_9 (x_A^3 \times x_B)$$

$$+w_{10} (x_B^3 \times x_C^3)$$

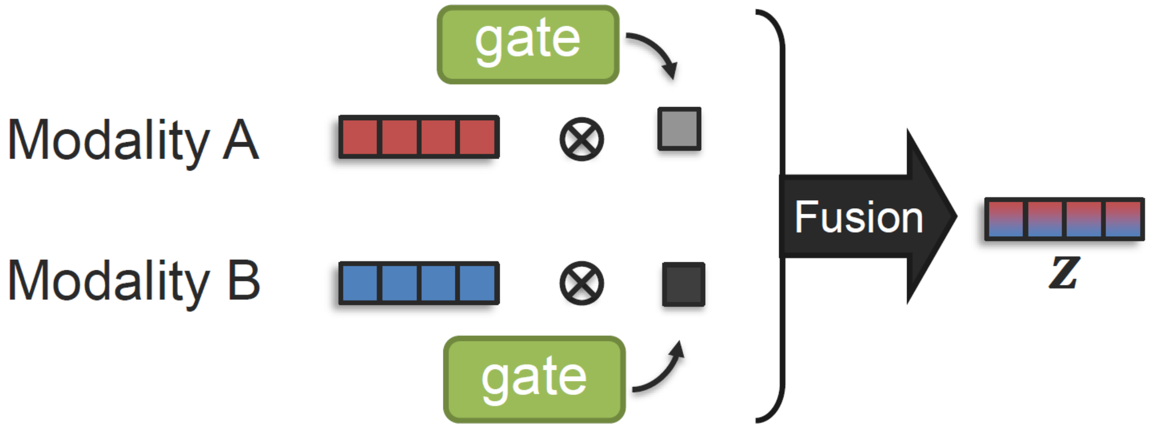
Gated Fusion



Example with additive fusion:

$$z = g_A(x_A, x_B) \cdot x_A + g_B(x_A, x_B) \cdot x_B$$

→ g_A and g_B can be seen as attention functions



→ Gating output can be one weight for the whole modality

Arevalo et al., Gated Multimodal Units for information fusion, ICLR-workshop 2017

Gating Module (aka, attention module)

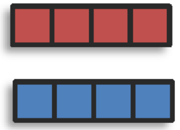
Input

What should it be?

Target modality

Other modality

All modality

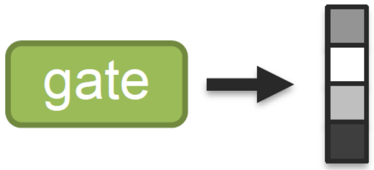


"Neural network designed to mask unwanted signal from propagating forward" (gating)

…or with a more positive view:

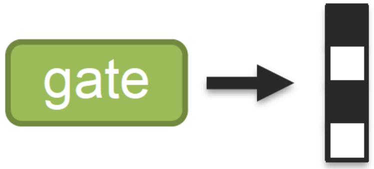
"Neural network designed to select preferable signal to move forward" (attention)

Soft attention



Easier to compute derivative (gradient)

Hard attention



Derivative is harder (e.g., use reinforcement learning)

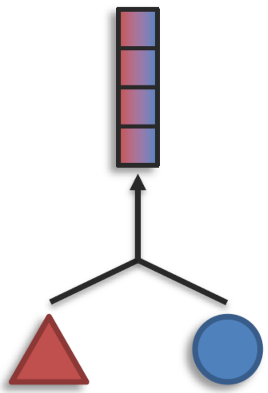
Chen et al., Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning, ICMI 2017

Task 1: Representation (表示)

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

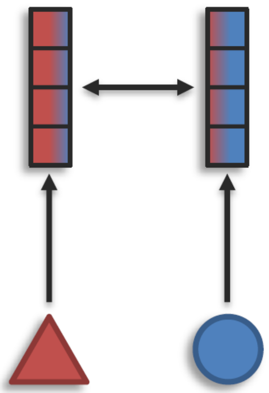
Sub-challenges:

Fusion



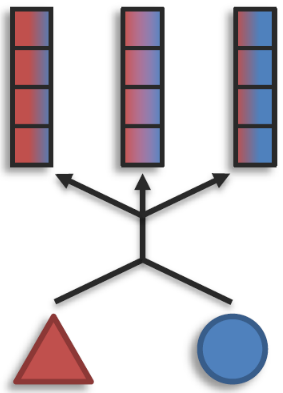
modalities **>** # representations

Coordination



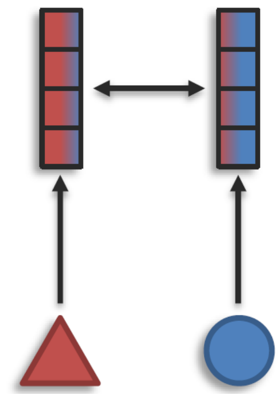
modalities **=** # representations

Fission



modalities **<** # representations

Sub-Challenge 1b: Representation Coordination

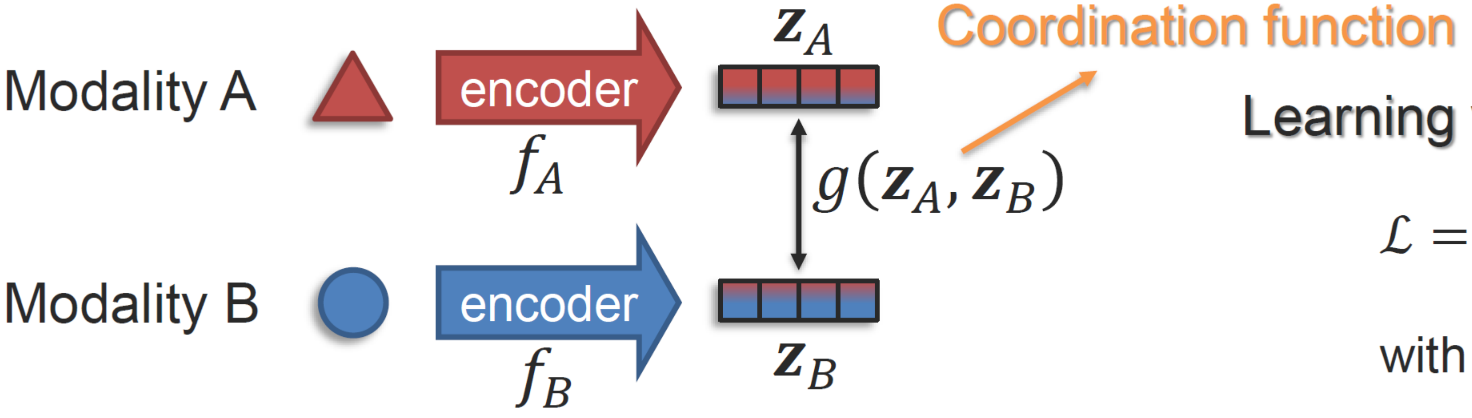


Definition: Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions

Strong Coordination:

Partial Coordination:

Coordination Function



Learning with coordination function:

$$\mathcal{L} = g(f_A(\blacktriangle), f_B(\bullet))$$

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

⇨ Requires paired data

Examples of coordination function:

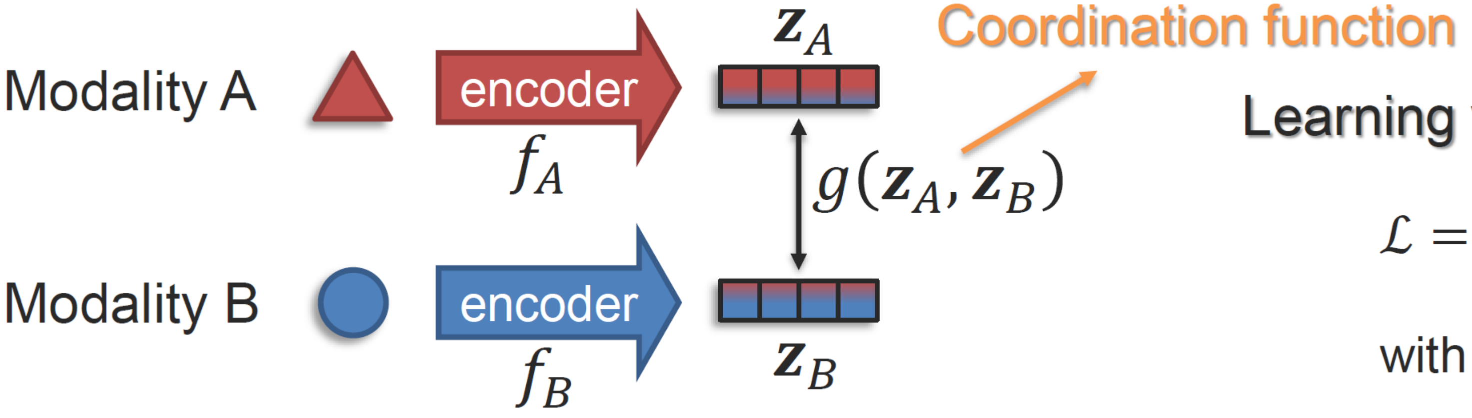
① Cosine similarity:

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

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:

$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

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

Requires paired data

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

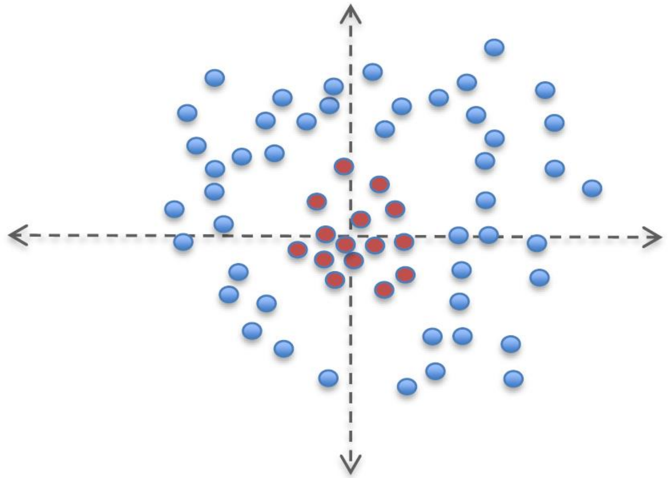
All these examples bring relatively strong coordination between modalities

Kernel Function

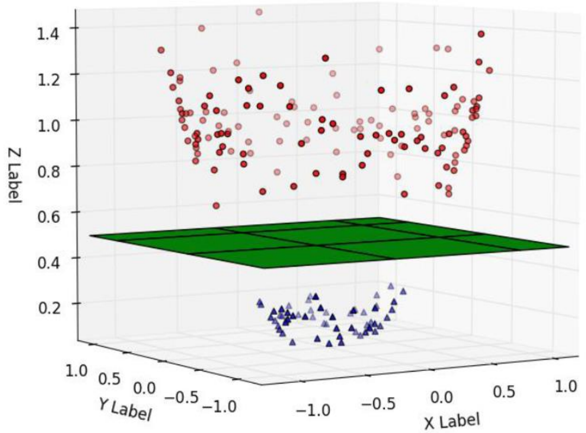
A kernel function: Acts as a similarity metric between data points

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

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



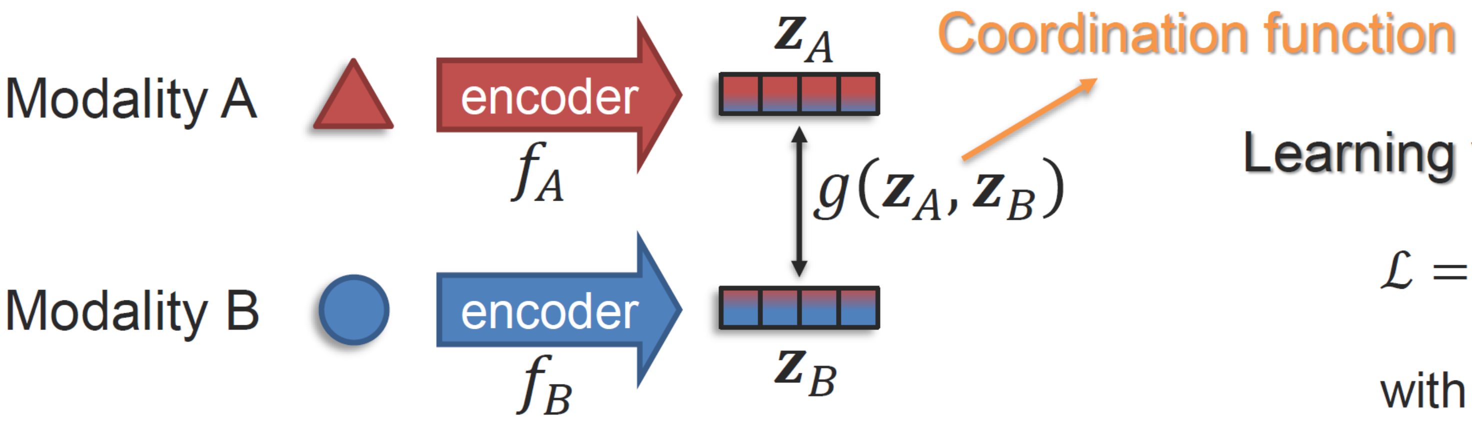
Not linearly separable in x space



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

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

Coordination Function



Learning with coordination function:

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

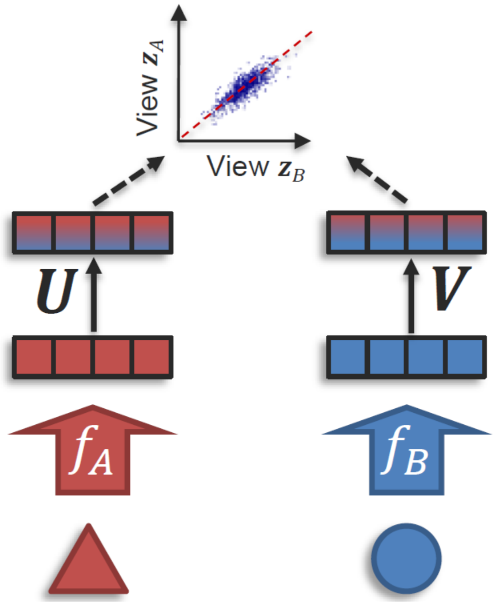
with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

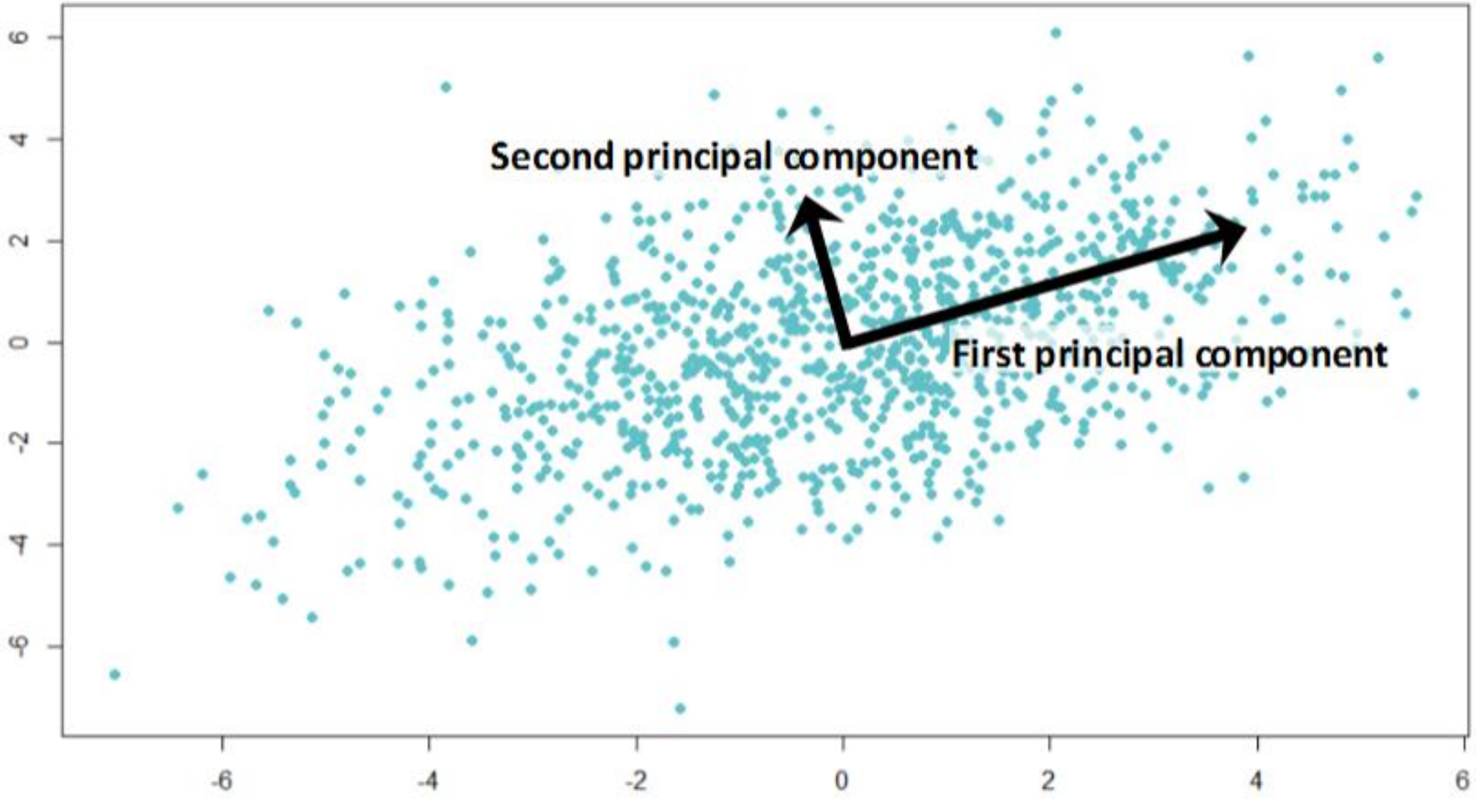
③ Canonical Correlation Analysis (CCA):

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

➡ CCA includes multiple projections, all orthogonal with each others

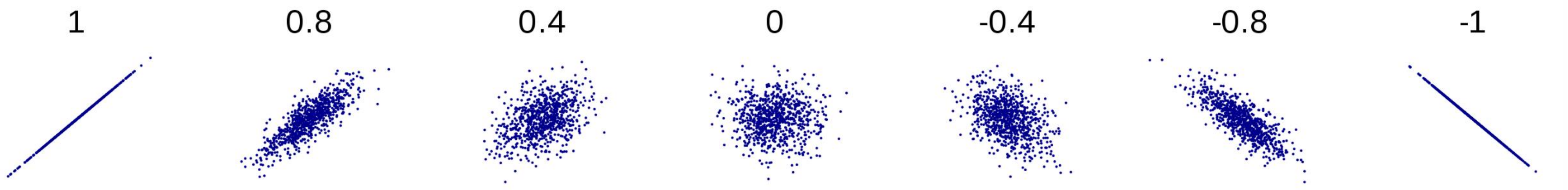


Retrospect: Principal Component Analysis (PCA)



$$\arg\max \ \|\boldsymbol{u}^\top \boldsymbol{X}\|^2$$
$$\text{s.t.}\ \boldsymbol{u}^\top \boldsymbol{u} = 1$$

Correlation



$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

$$X, Y \text{ independent} \quad \Rightarrow \quad \rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated})$$

$$\rho_{X,Y} = 0 \quad (X, Y \text{ uncorrelated}) \quad \not\Rightarrow \quad X, Y \text{ independent}$$

Correlation

1. 二次函数关系的例子

设随机变量 X 均匀分布在区间 $[-1,1]$，定义 $Y=X^2$。虽然 X 和 Y 之间显然不是独立的，因为 Y 完全由 X 决定，但它们是 **不相关的**，因为协方差为零。

- **证明不相关性**:

$$\text{Cov}(X,Y)=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$$

由于 $Y=X^2$，我们可以计算:

$$\mathbb{E}[X]=0 \quad (\text{因为 } X \text{ 是均匀分布的，对称性导致均值为 } 0)$$

$$\mathbb{E}[X^3]=0 \quad (\text{立方对称分布于 [-1, 1]，故期望为 } 0)$$

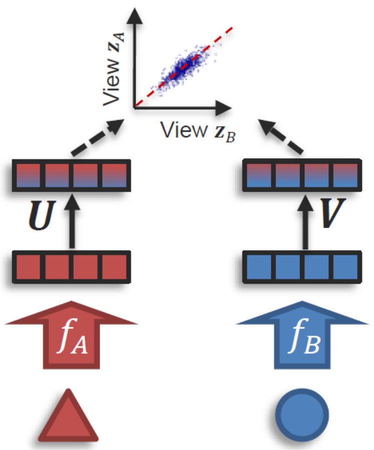
因此，$\text{Cov}(X,Y)=0$，它们不相关，但显然 X 和 Y 不是独立的。

https://chatgpt.com/

Correlated Projection

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

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \arg\max \frac{\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XY} \boldsymbol{v}}{\sqrt{\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XX} \boldsymbol{u}} \sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}_{YY} \boldsymbol{v}}}$$



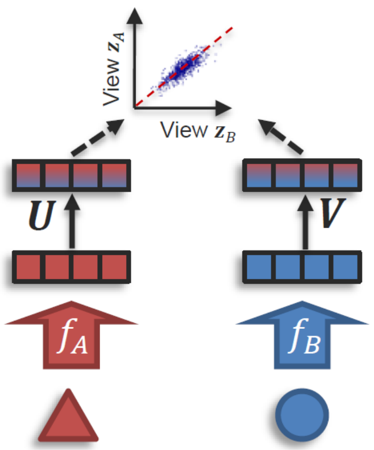
Two views X, Y where same instances have the same color

Remember that X and Y consist of paired data

Correlated Projection

The **first pair of canonical variables:**

$$(\boldsymbol{u}_1, \boldsymbol{v}_1) = \arg\max \frac{\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XY} \boldsymbol{v}}{\sqrt{\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XX} \boldsymbol{u}} \sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}_{YY} \boldsymbol{v}}}$$



2

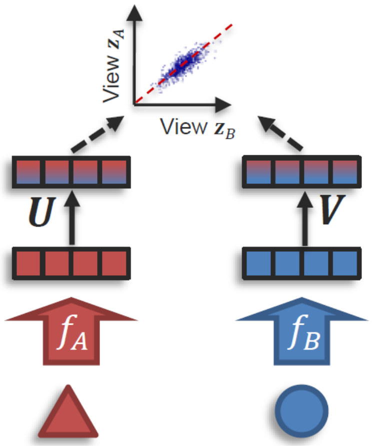
Since this objective function is invariant to scaling, we can constraint the projections to have unit variance:

$$\boldsymbol{u}_1^\top \boldsymbol{\Sigma}_{XX} \boldsymbol{u}_1 = \boldsymbol{v}_1^\top \boldsymbol{\Sigma}_{YY} \boldsymbol{v}_1 = \mathbf{1}$$

Correlated Projection

The **k-th pair of canonical variables:**

$$(\boldsymbol{u}_k, \boldsymbol{v}_k) = \arg\max \frac{\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XY} \boldsymbol{v}}{\sqrt{\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XX} \boldsymbol{u}} \sqrt{\boldsymbol{v}^\top \boldsymbol{\Sigma}_{YY} \boldsymbol{v}}}$$

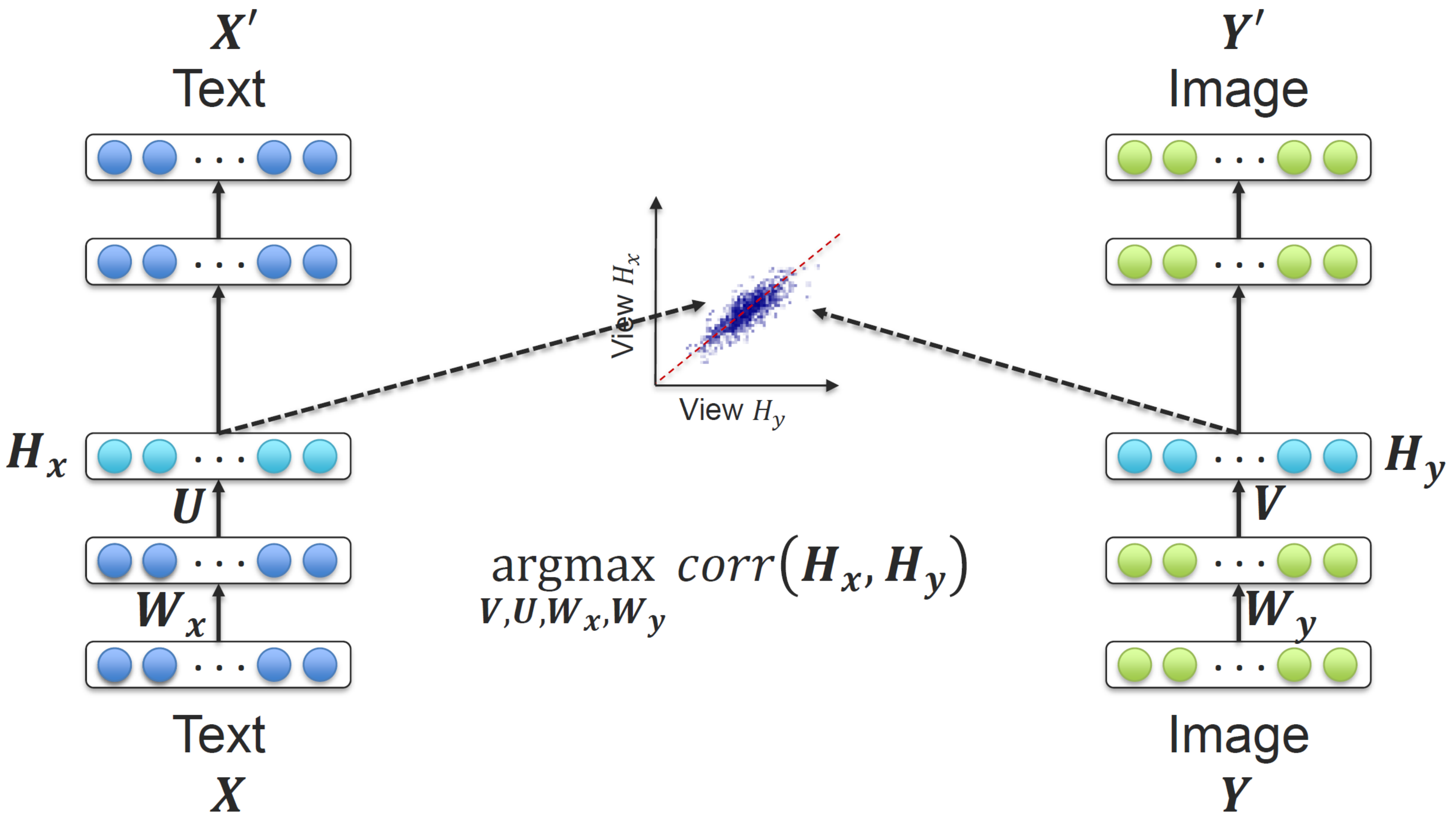


3. We want these multiple projection pairs to be orthogonal ("canonical") to each other:

$$\boldsymbol{u}^\top \boldsymbol{\Sigma}_{XX} \boldsymbol{u}_j = \boldsymbol{v}^\top \boldsymbol{\Sigma}_{YY} \boldsymbol{v}_j = \boldsymbol{0}, \forall j = 1, \cdots, k-1$$

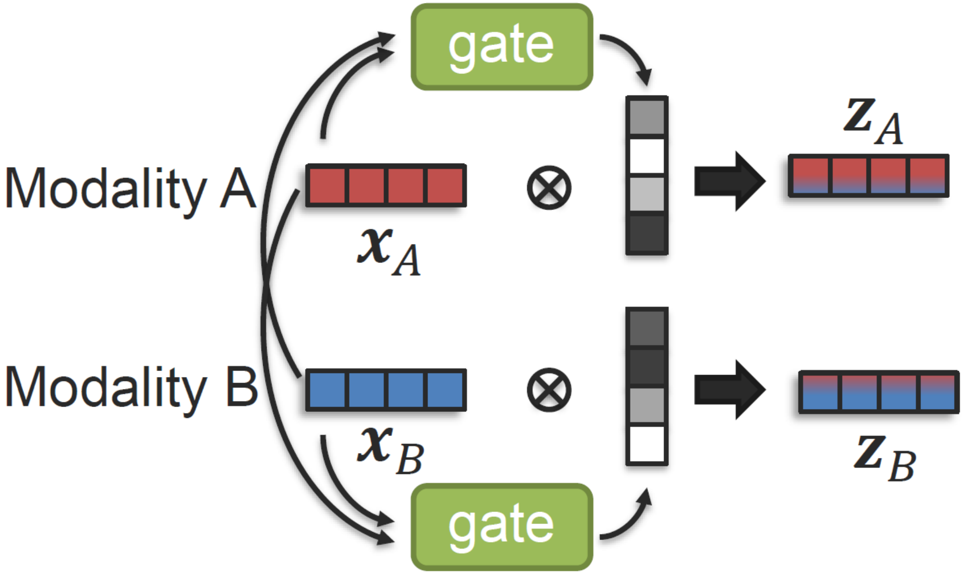
https://en.wikipedia.org/wiki/Canonical_correlation

Deep Canonically Correlated Autoencoders (DCCAE)



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

Gated Coordination



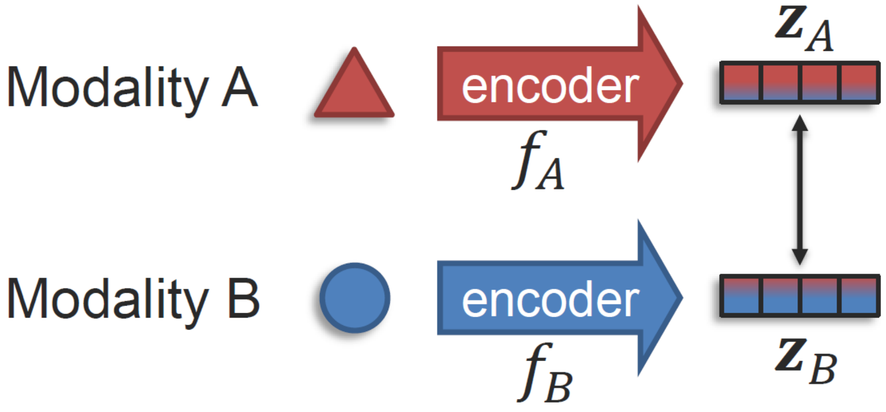
Gated coordination:

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

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

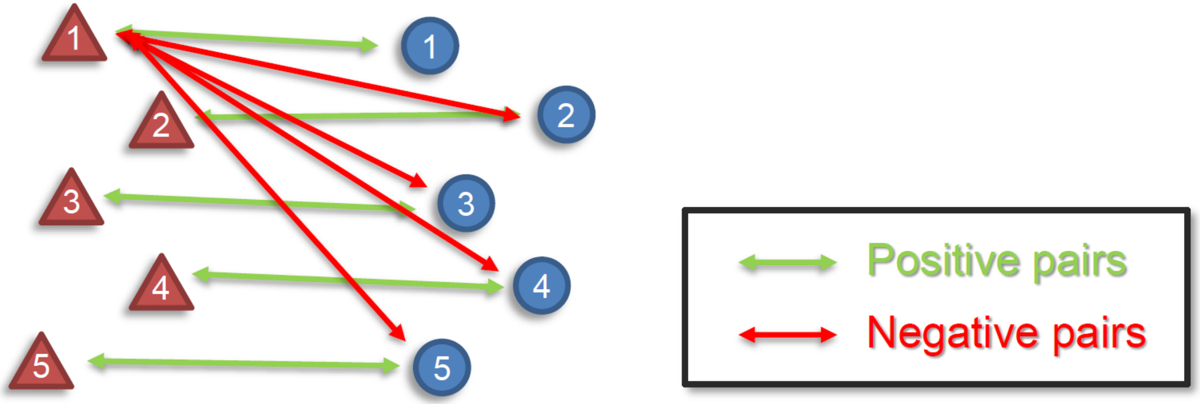
Related to attention modules in transformers

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

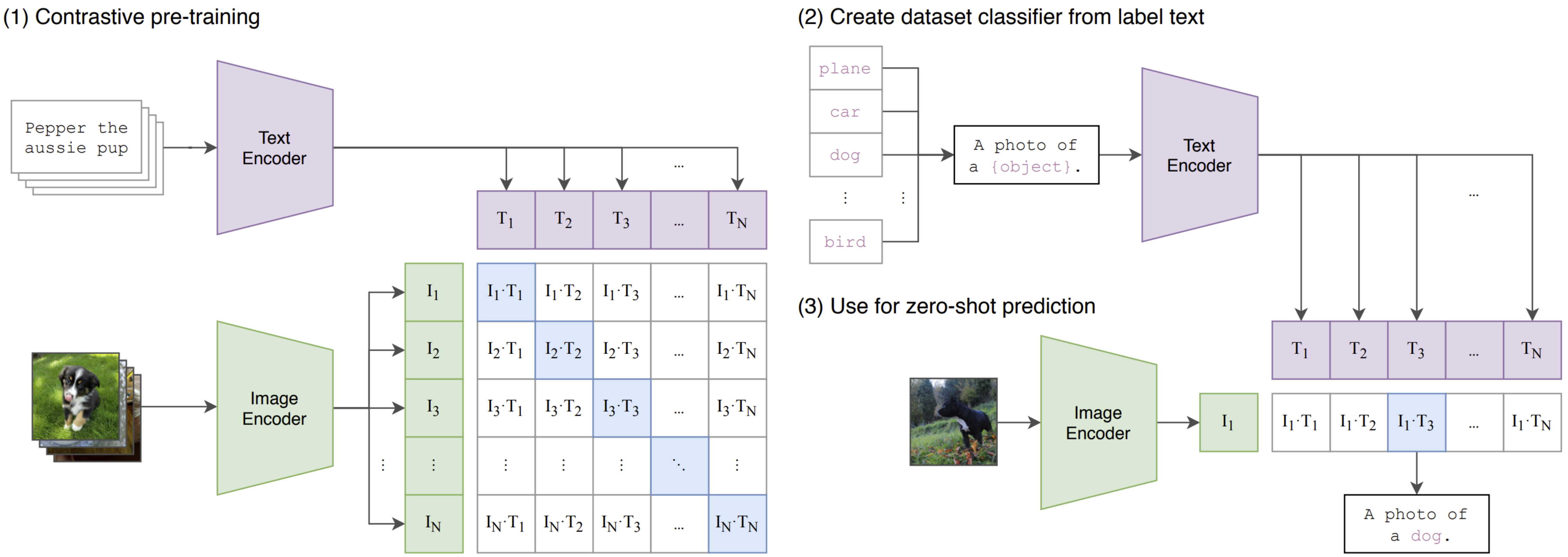
negative pair

Similarity functions are often cosine similarity

➡ Similar to hinge loss

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

Zero-Shot Image Classification



Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 （**Citations： 9211-> 22407** ）

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

Pretrained Dataset (not open-sourved by openAI)

we constructed a new dataset of **400 million (image, text) pairs** collected form a variety of publicly available sources on the Internet. To attempt to cover as broad a set of visual concepts as possible, we search for (image, text) pairs as part of the construction process whose text includes one of a set of 500,000 queries. We approximately class balance the results by including up to 20,000 (image, text) pairs per query. The resulting dataset has a similar total word count as the WebText dataset used to train GPT-2. We refer to this dataset as **WIT** for **WebImageText**.

Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 （**Citations： 9211-> 22407** ）

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

Encoders

Model	Learning rate	Embedding dimension	Input resolution	ResNet blocks	width	Text Transformer layers	width	heads
RN50	5×10^{-4}	1024	224	(3, 4, 6, 3)	2048	12	512	8
RN101	5×10^{-4}	512	224	(3, 4, 23, 3)	2048	12	512	8
RN50x4	5×10^{-4}	640	288	(4, 6, 10, 6)	2560	12	640	10
RN50x16	4×10^{-4}	768	384	(6, 8, 18, 8)	3072	12	768	12
RN50x64	3.6×10^{-4}	1024	448	(3, 15, 36, 10)	4096	12	1024	16

Table 19. CLIP-ResNet hyperparameters

Model	Learning rate	Embedding dimension	Input resolution	Vision Transformer layers	width	heads	Text Transformer layers	width	heads
ViT-B/32	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-B/16	5×10^{-4}	512	224	12	768	12	12	512	8
ViT-L/14	4×10^{-4}	768	224	24	1024	16	12	768	12
ViT-L/14-336px	2×10^{-5}	768	336	24	1024	16	12	768	12

Table 20. CLIP-ViT hyperparameters

Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 （**Citations： 9211-> 22407** ）

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

```
# image_encoder - ResNet or Vision Transformer
# text_encoder  - CBOW or Text Transformer
# I[n, h, w, c] - minibatch of aligned images
# T[n, l]       - minibatch of aligned texts
# W_i[d_i, d_e] - learned proj of image to embed
# W_t[d_t, d_e] - learned proj of text to embed
# t             - learned temperature parameter

# extract feature representations of each modality
I_f = image_encoder(I) #[n, d_i]
T_f = text_encoder(T)  #[n, d_t]

# joint multimodal embedding [n, d_e]
I_e = l2_normalize(np.dot(I_f, W_i), axis=1)
T_e = l2_normalize(np.dot(T_f, W_t), axis=1)

# scaled pairwise cosine similarities [n, n]
logits = np.dot(I_e, T_e.T) * np.exp(t)

# symmetric loss function
labels = np.arange(n)
loss_i = cross_entropy_loss(logits, labels, axis=0)
loss_t = cross_entropy_loss(logits, labels, axis=1)
loss   = (loss_i + loss_t)/2
```

Symmetric InfoNCE (Noise Contrastive Estimation) loss

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\text{image}} + \mathcal{L}_{\text{text}}\right)$$

Positive pair

$$\mathcal{L}_{\text{image}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\text{sim}\left(\boldsymbol{z}_{i,\text{image}}, \boldsymbol{z}_{i,\text{text}}\right)}{\sum_{j=1}^{N} \text{sim}\left(\boldsymbol{z}_{i,\text{image}}, \boldsymbol{z}_{j,\text{text}}\right)}$$

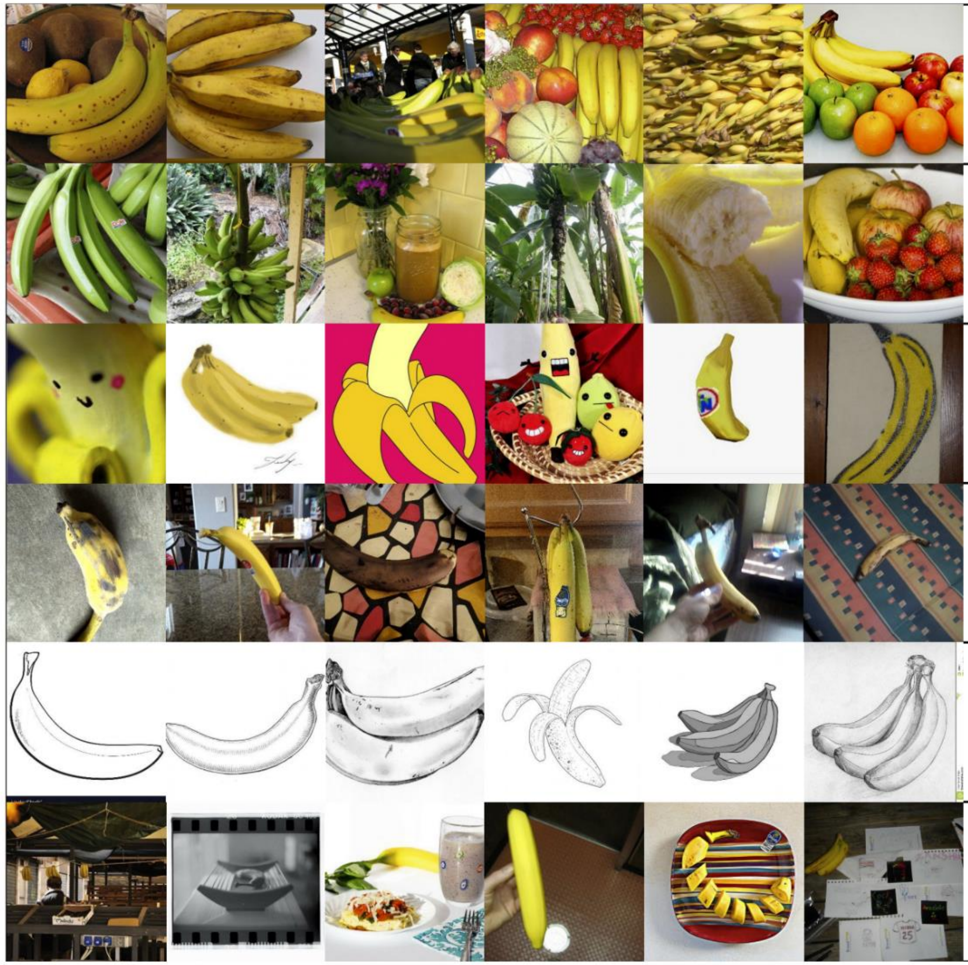
Negative pairs

$$\mathcal{L}_{\text{text}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\text{sim}\left(\boldsymbol{z}_{i,\text{image}}, \boldsymbol{z}_{i,\text{text}}\right)}{\sum_{j=1}^{N} \text{sim}\left(\boldsymbol{z}_{j,\text{image}}, \boldsymbol{z}_{i,\text{text}}\right)}$$

Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 （**Citations： 9211-> 22407** ）

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

Dataset Examples

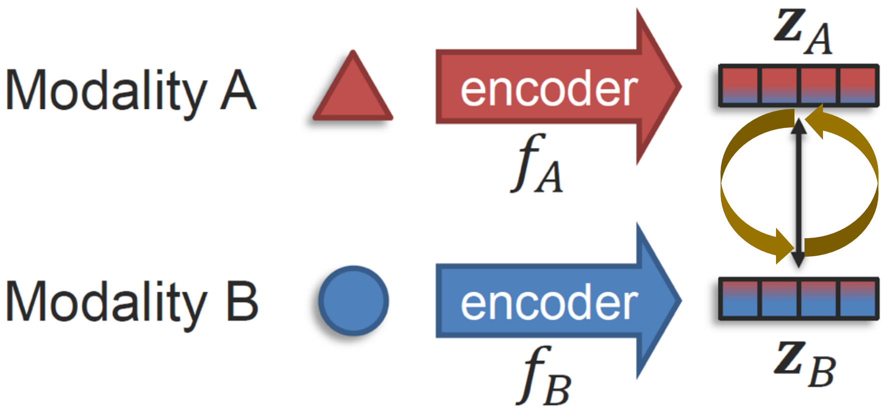


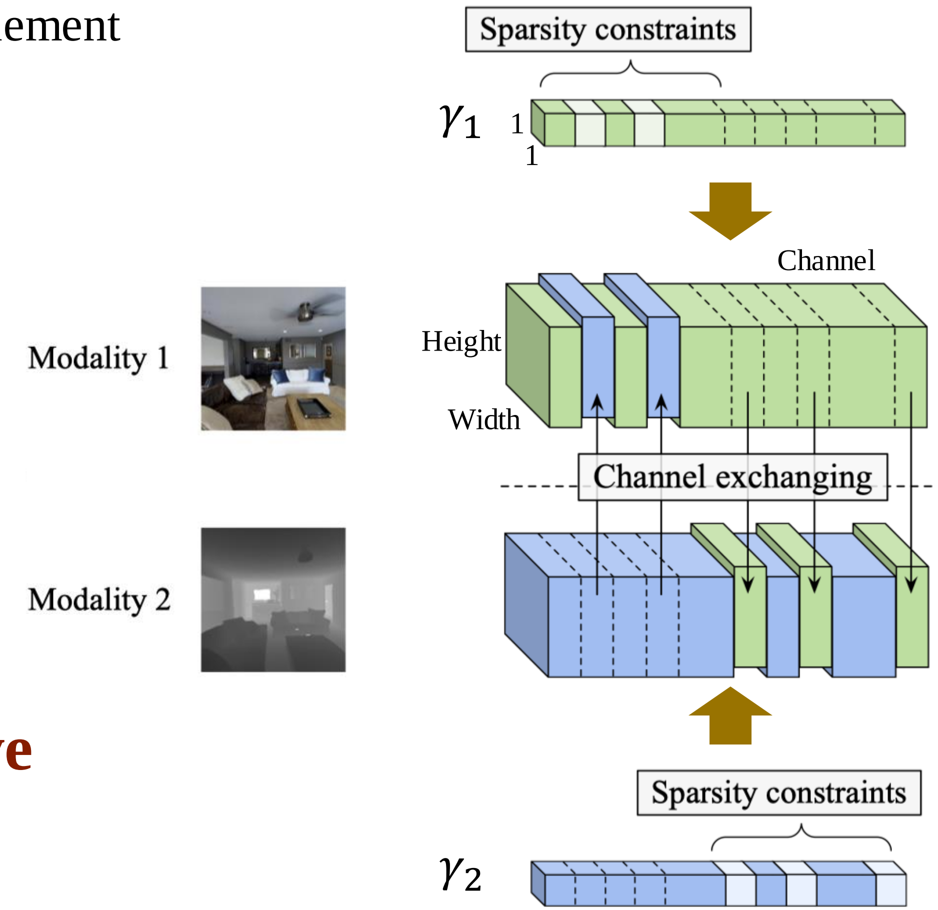
	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet	**76.2**	**76.2**	0%
ImageNetV2	64.3	**70.1**	+5.8%
ImageNet-R	37.7	**88.9**	+51.2%
ObjectNet	32.6	**72.3**	+39.7%
ImageNet Sketch	25.2	**60.2**	+35.0%
ImageNet-A	2.7	**77.1**	+74.4%

Learning Transferable Visual Models From Natural Language Supervision, ICML 2021 （**Citations： 9211-> 22407** ）

Homogeneous Coordination with Channel Exchanging

Homogeneous Multimodal Learning: The modalities to fuse are of the same shape; there is certain correspondence between their each element





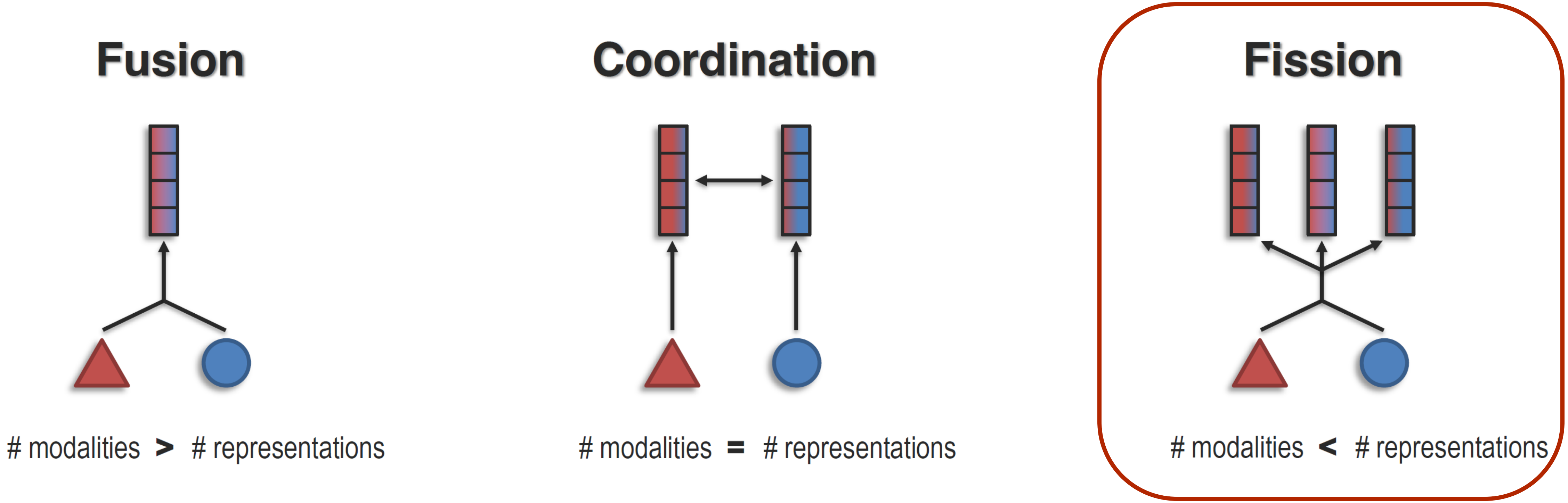
Parameter-free, **Self-adaptive**

Deep Multimodal Fusion by Channel Exchanging, NeurIPS 2020.

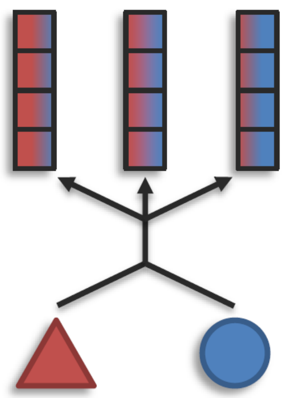
Task 1: Representation (表示)

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

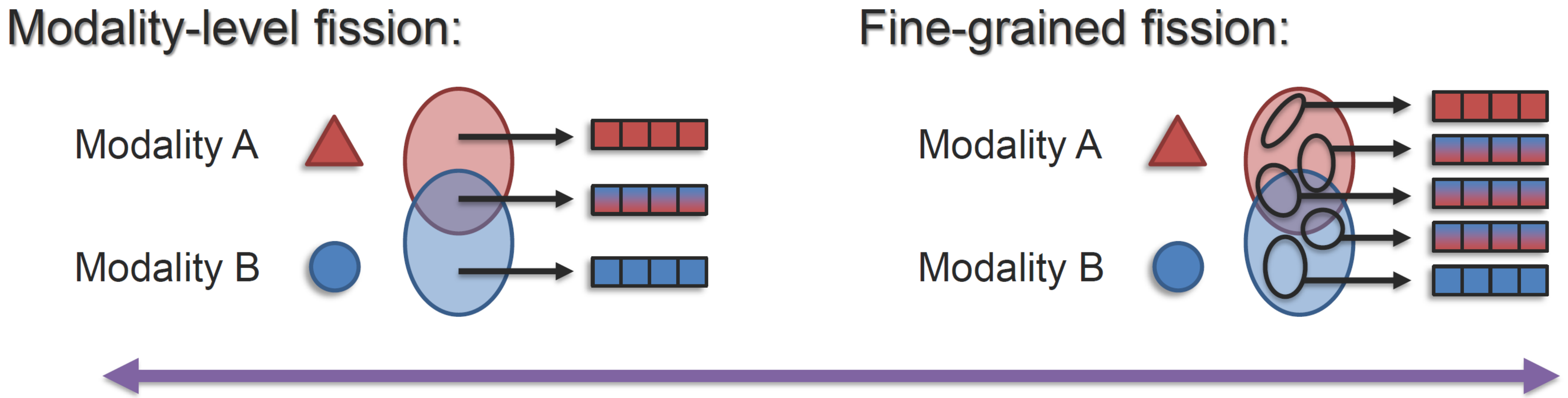
Sub-challenges:



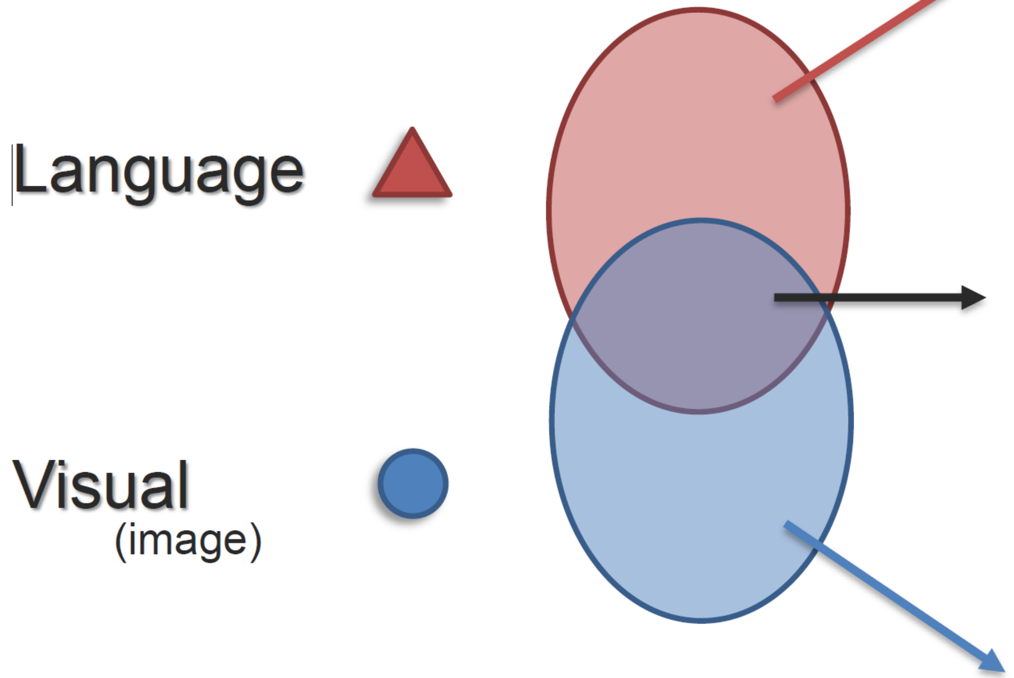
Sub-Challenge 1c: Representation Fission



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



Modality-Level Fission



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

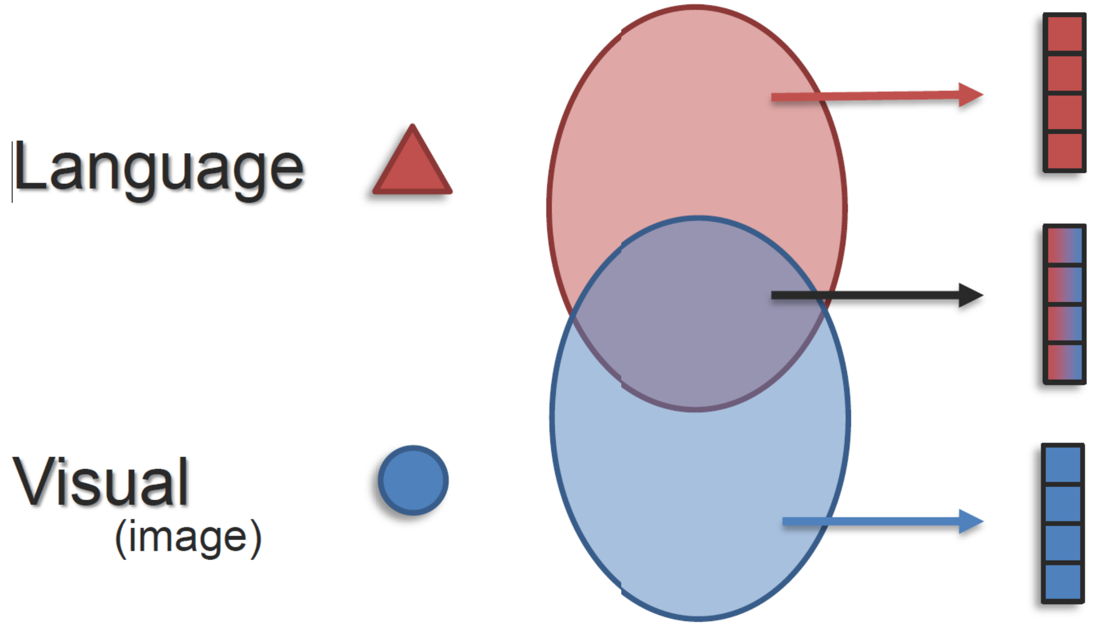
Information in both modalities

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

Information primarily in visual modality

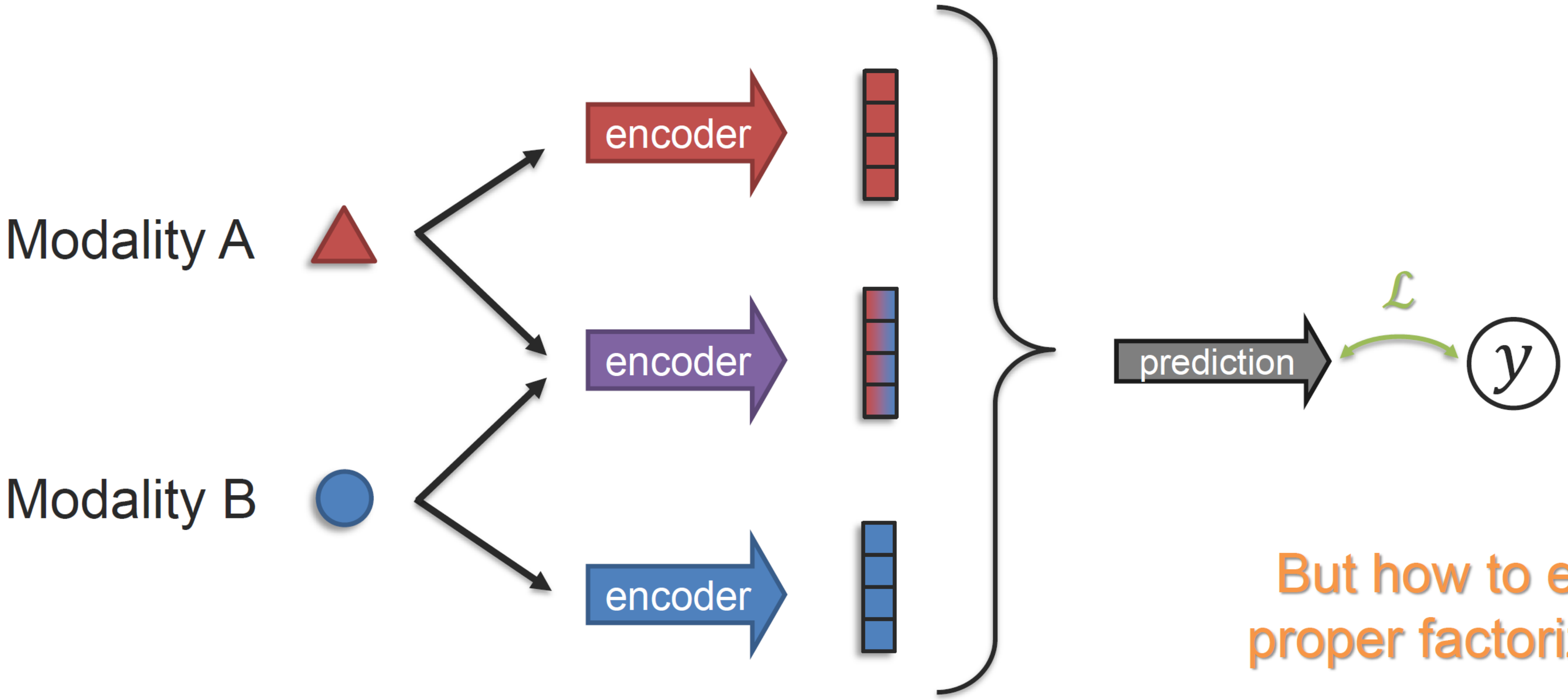
- Texture, visual appearance
- Depth, perspective, motion
- …

Modality-Level Fission



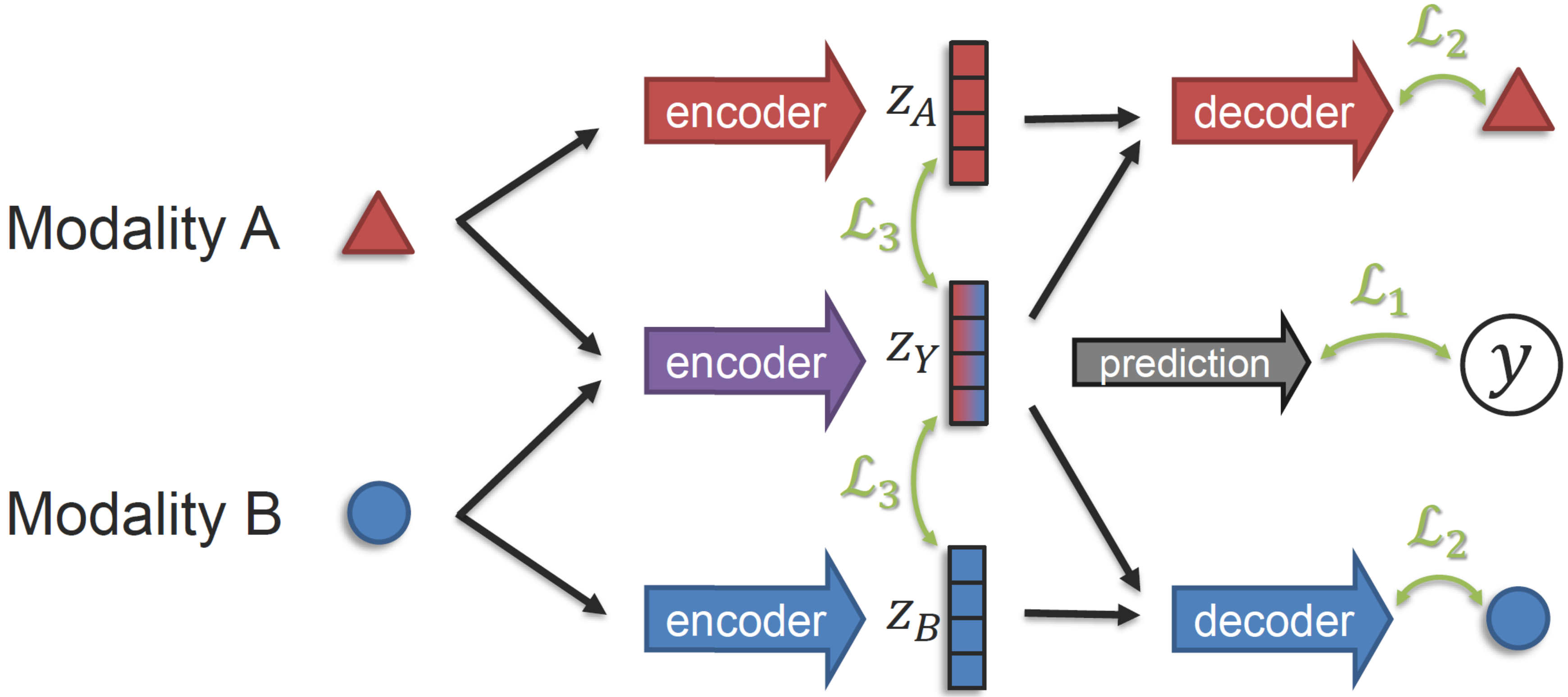
How to learn factorized multimodal representations?

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

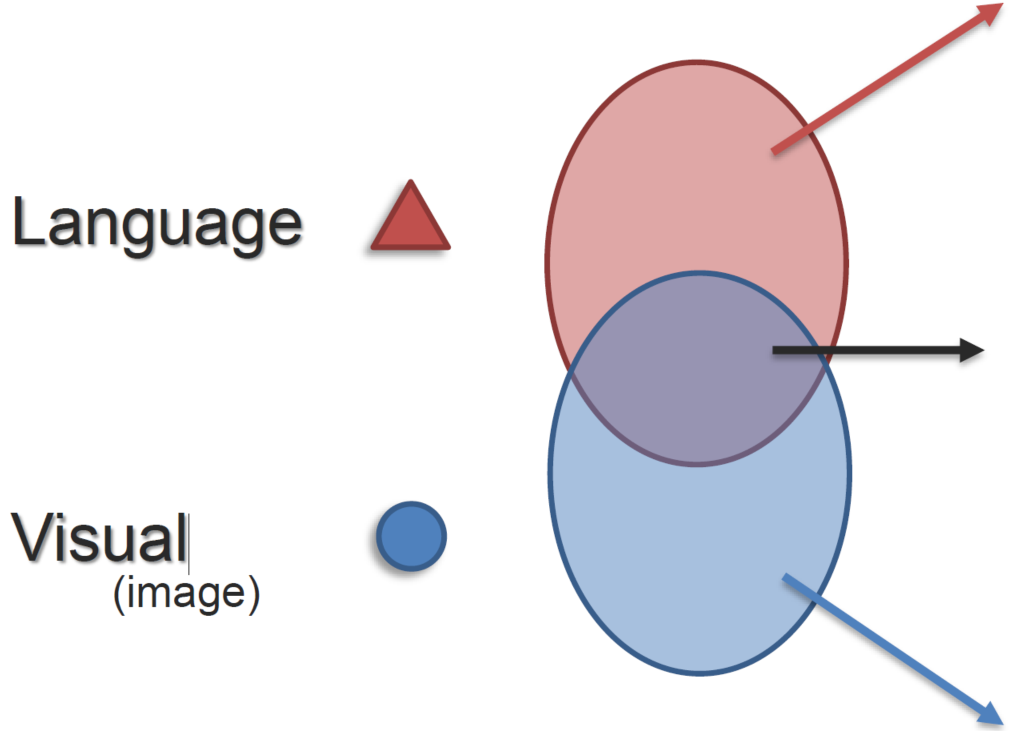
$\mathcal{L}_2$: generative

$\mathcal{L}_3$: 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



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

Information in both modalities

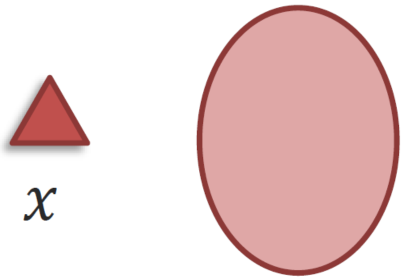
- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality

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

Information and Entropy – Information Theory

Language



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 $I(x)$

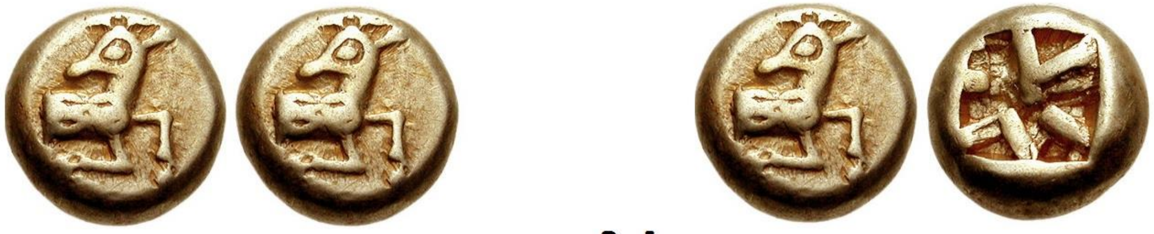
$$I(x) \sim \frac{1}{p(x)}$$

➡ 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



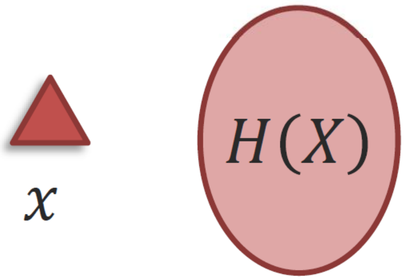
$$S = \log_2 2^N = N = 2$$

Information entropy (in bits) is the log-base-2 of the number of possible outcomes. With two coins there are four outcomes HH-HT-TH-TT, and the entropy is two bits.

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 $\mathcal{X}$

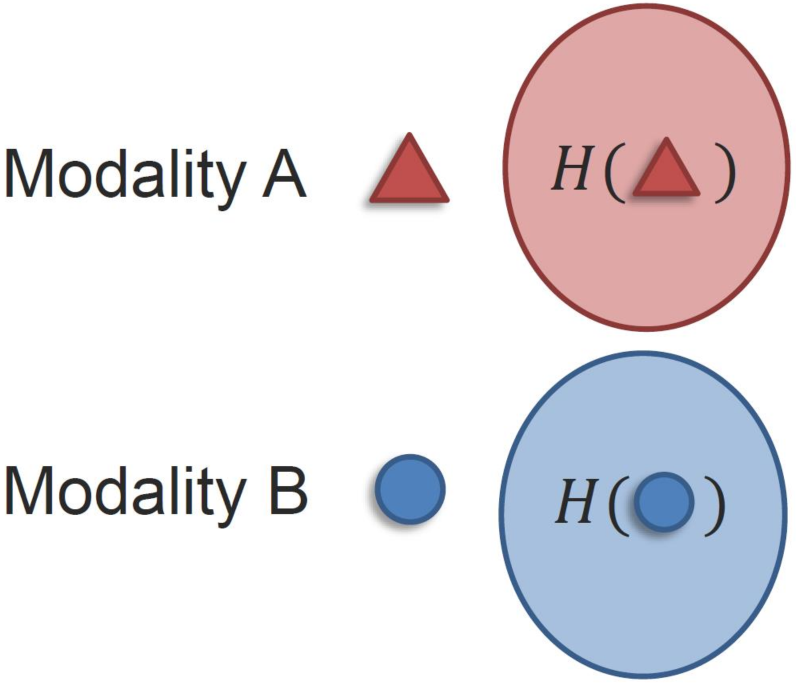
$$H(X) = \mathbb{E}[I(X)] = \mathbb{E}[-\log(p(X))] = -\sum_{x \in \mathcal{X}} p(X)\log(p(X))$$

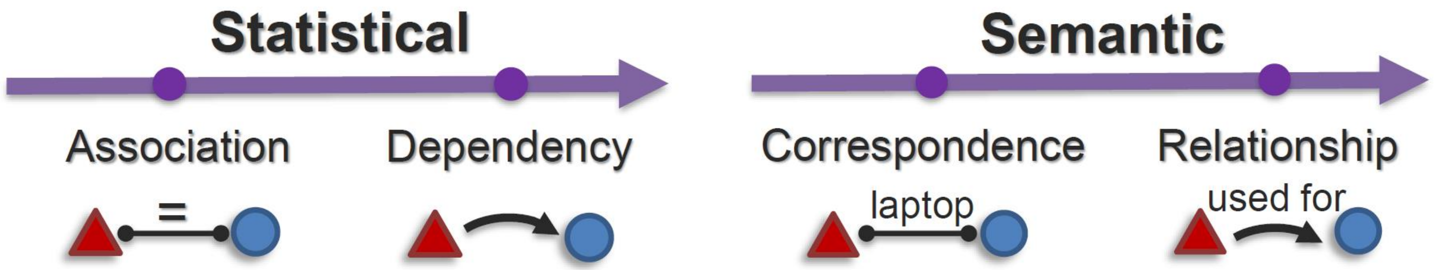
- Entropy can also be defined for continuous random variables

Information and Entropy

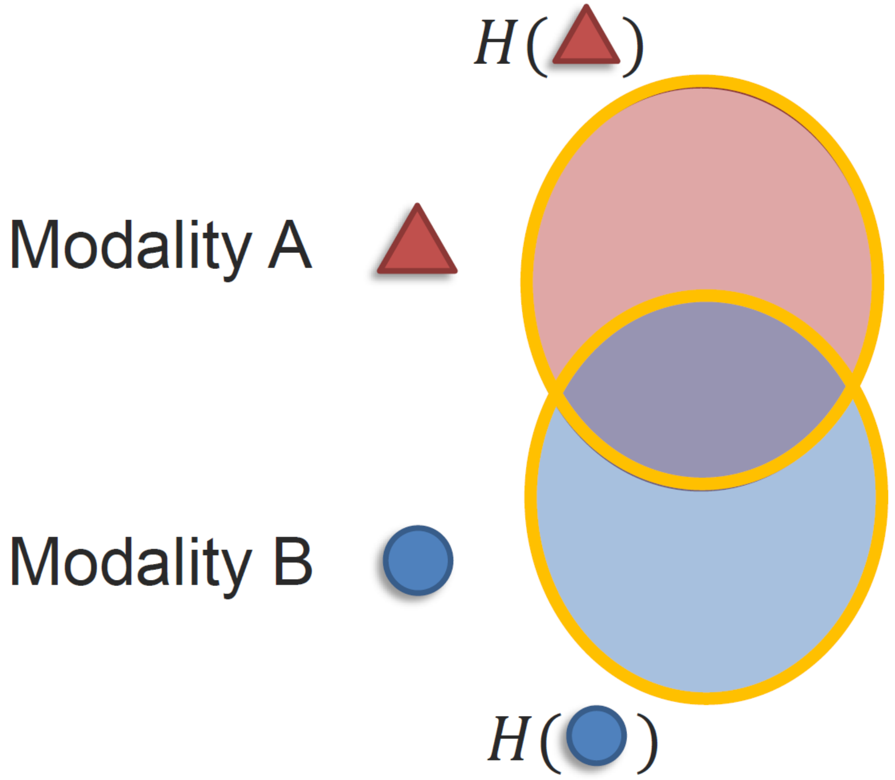
If no overlapping information

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

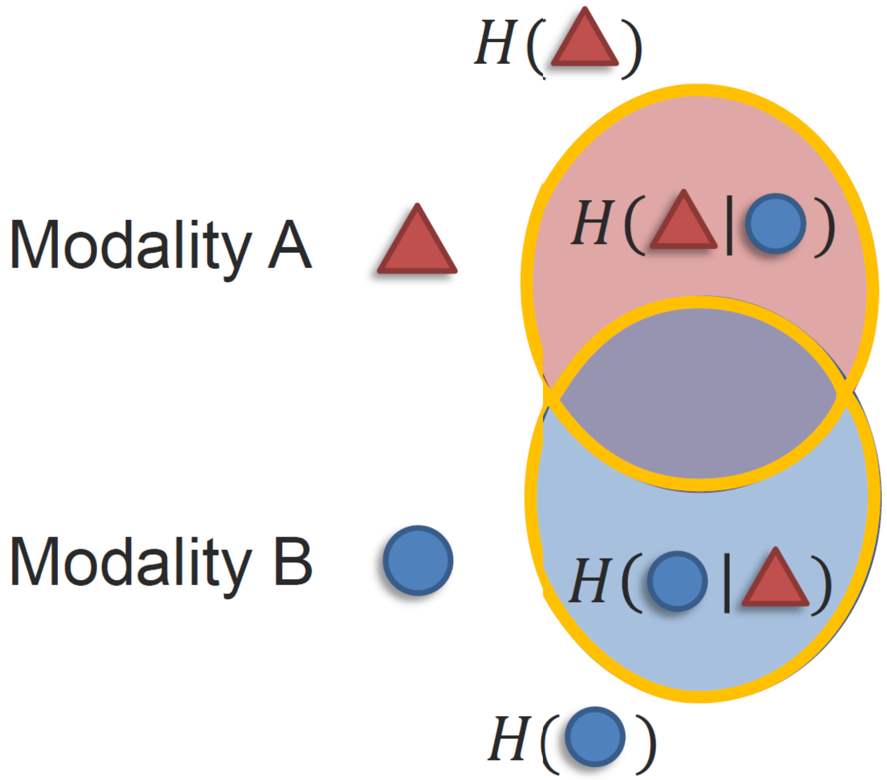




Entropy with Two Modalities



Entropy with Two Modalities

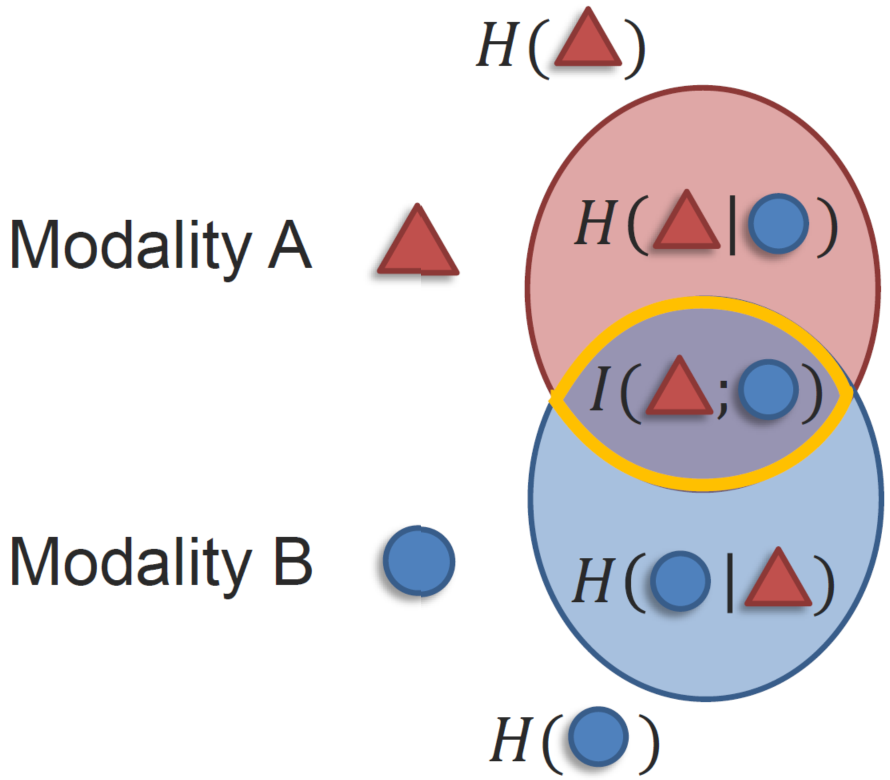


Conditional entropy $H(Y|X)$

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

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

Entropy with Two Modalities



Mutual information $I(X;Y)$

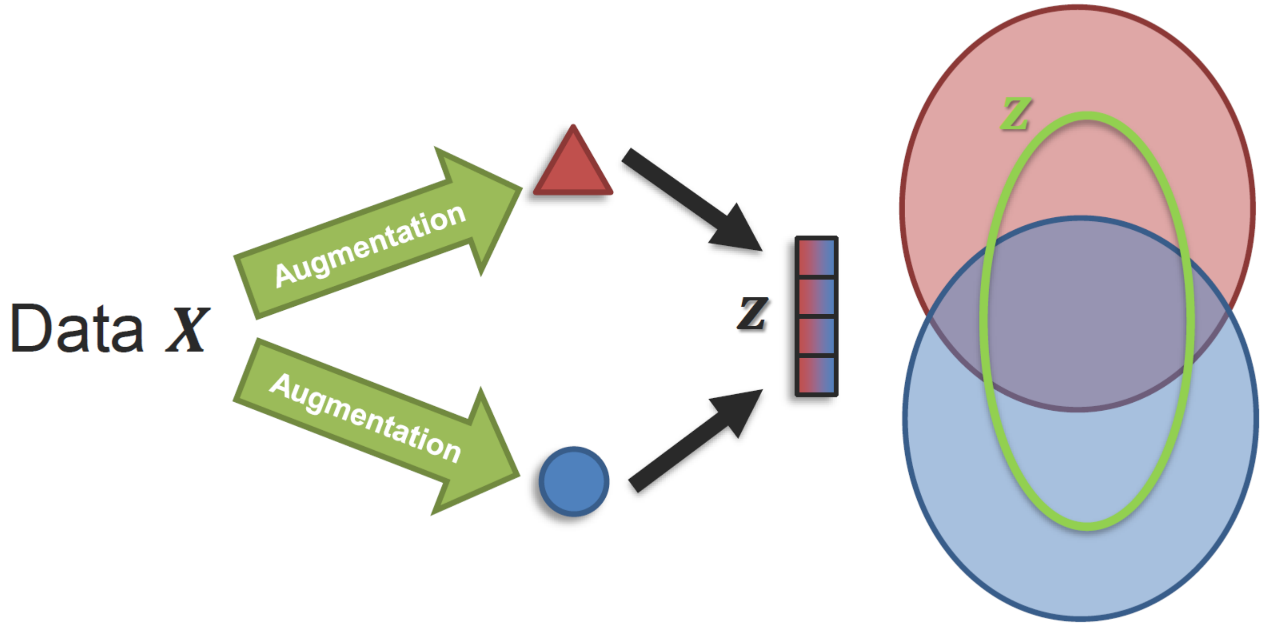
$$I(X;Y) = H(X) - H(X|Y)$$

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

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

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

Link with Self-Supervised Learning



① Maximize the mutual information

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

➡ Related to contrastive learning

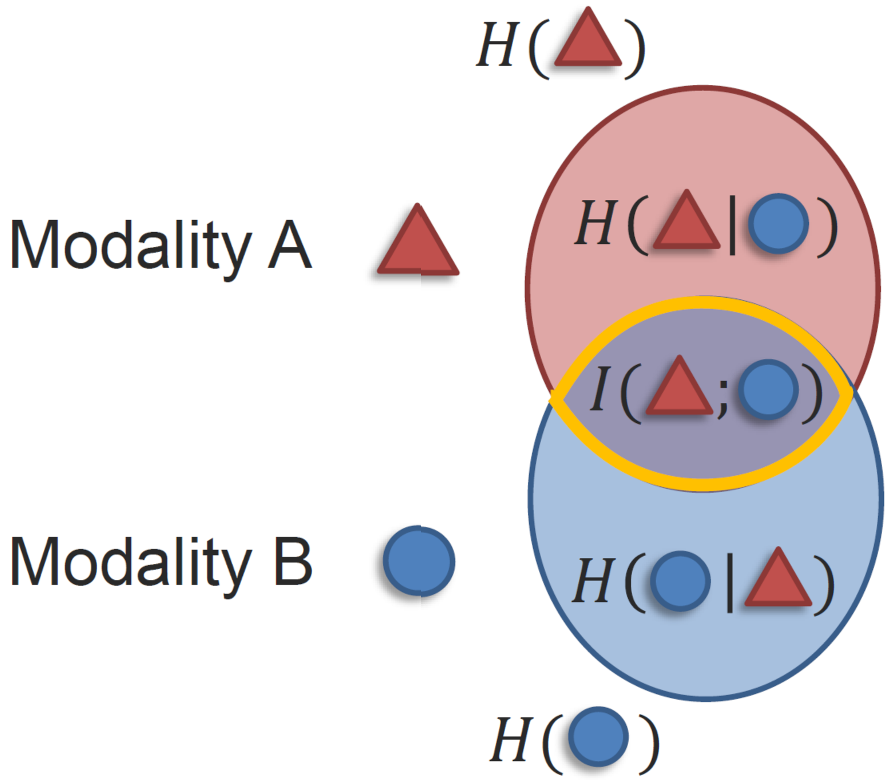
② Minimize the conditional entropy

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

Information theory gives us a path towards disentangled representation learning

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

Some facts about information theory



1. Properties of mutual information:

$$I(X;Y) \geq 0, I(X;Y) = I(Y;X)$$

2. Subadditivity:

$$H(X) + H(Y) = H(X,Y) + I(X;Y) \geq H(X,Y)$$

$$H(X) = H(X|Y) + I(X;Y) \geq H(X|Y)$$

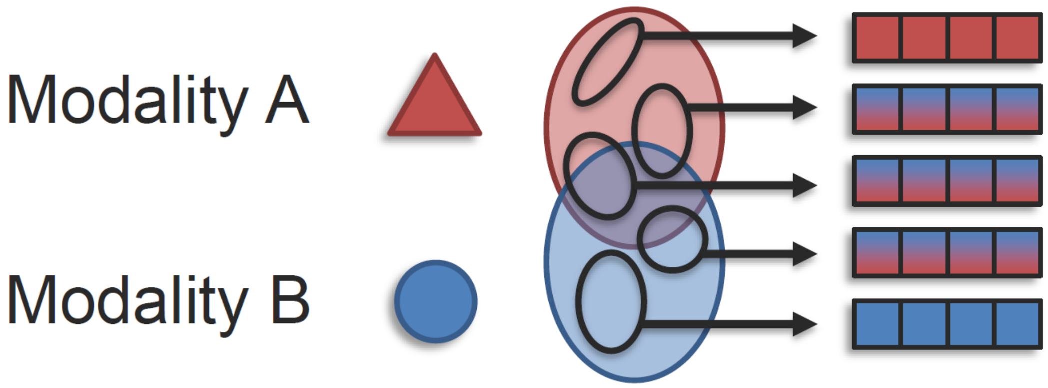
3. The entropy or the amount of information revealed by evaluating X and Y simultaneously is equal to: first evaluating the value of Y, then revealing the value of X given that you know the value of Y.

$$H(X,Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

But, do we have $H(X|Y) \geq 0$? ? ?

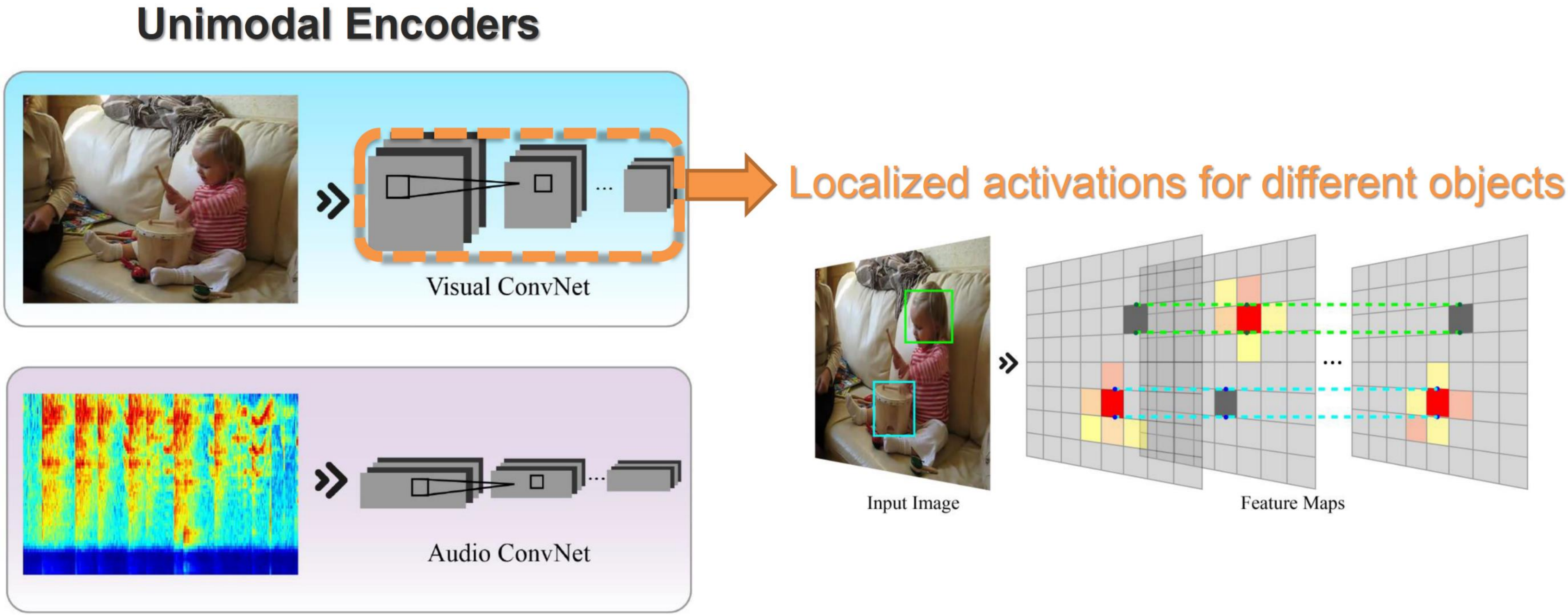
https://en.wikipedia.org/wiki/Conditional_entropy

Fine-Grained Fission



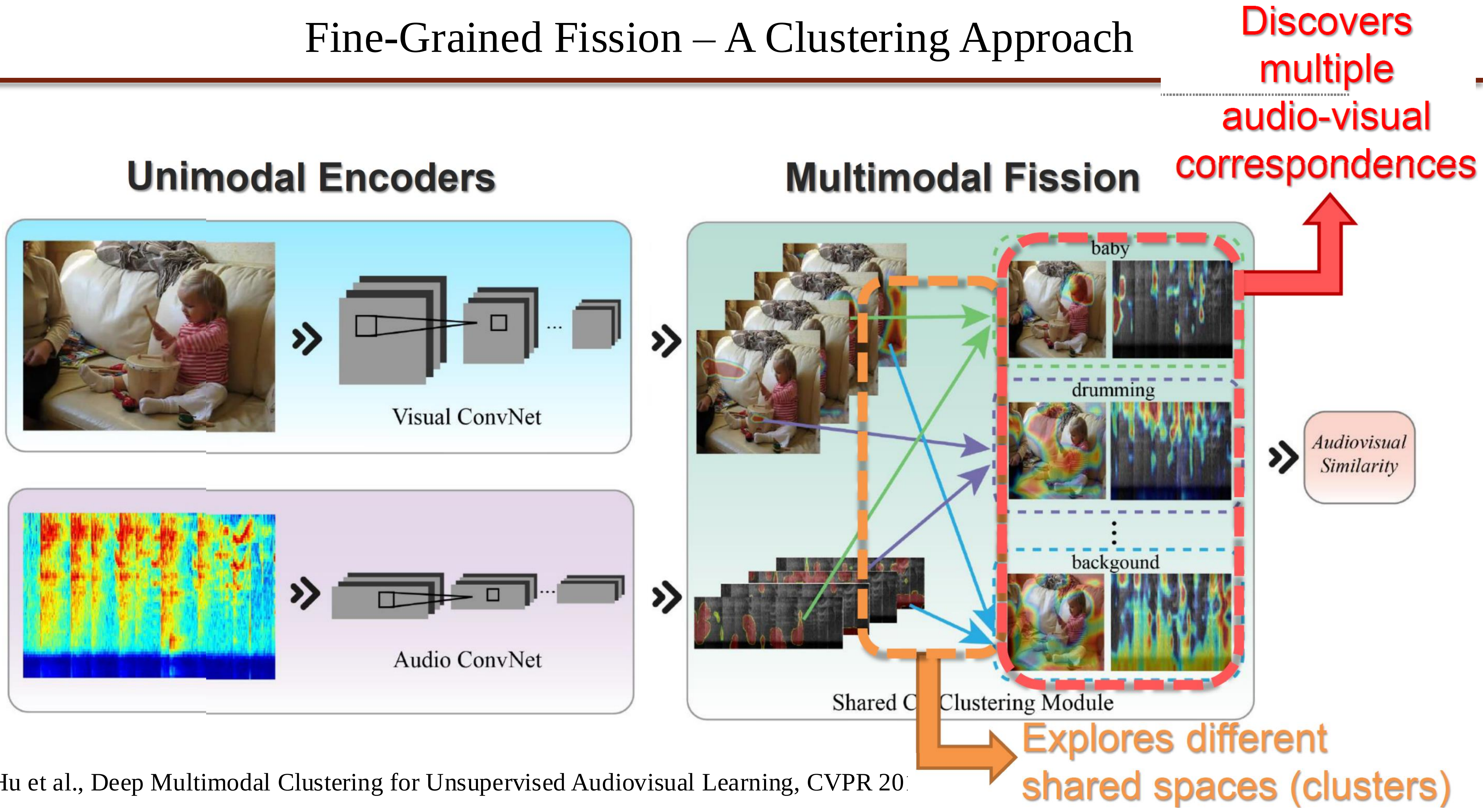
How to automatically discover these internal clusters, factors?

Fine-Grained Fission – A Clustering Approach



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

Fine-Grained Fission – A Clustering Approach



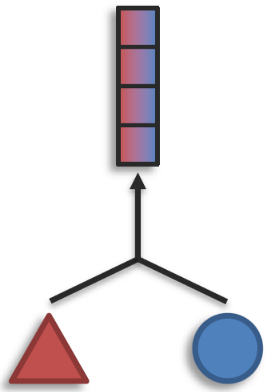
Hu et al., Deep Multimodal Clustering for Unsupervised Audiovisual Learning, CVPR 20

Task 1: Representation (表示)

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

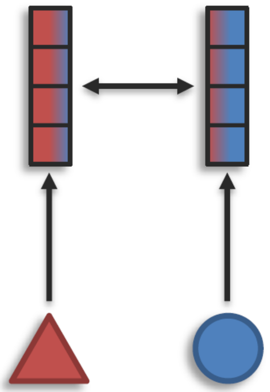
Sub-challenges:

Fusion



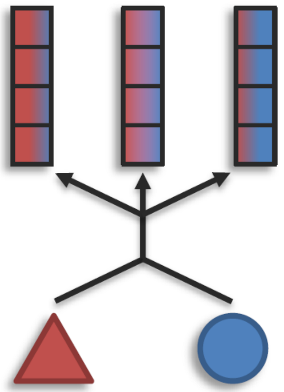
modalities **>** # representations

Coordination



modalities **=** # representations

Fission



modalities **<** # representations