Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 3.2: Multimodal Coordination and Fission

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Upcoming Course Assignments

Week 3 Reading Assignment (due Friday 9/11 at 8pm ET)

- Two papers: a first about unimodal and a second about multimodal

Proposal, aka First Assignment (due Sunday 9/20 at 8pm ET)

- Literature review of your research problem
- Second revision of your research ideas (i.e., top-down approach)

Coming soon: Proposal presentations

- No lectures on Tuesday 9/22 and Thursday 9/24
- Each team will schedule a 15mins timeslot
- 5 mins presentations + 10 mins discussion

Attendance using Poll Everywhere

1. Scan QR Code
   or
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

✗ *Do not try to sign-in*

3. Click ”Share location”

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 3.2: Multimodal Coordination and Fission

Louis-Philippe Morency

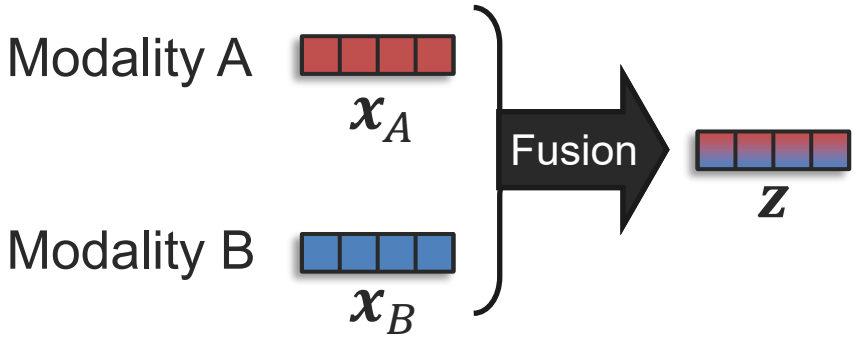
** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Objectives of today's class

- Representation fusion
  - Additive and multiplicative fusion
  - Gated fusion
- 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

Representation Fusion

Additive Fusion

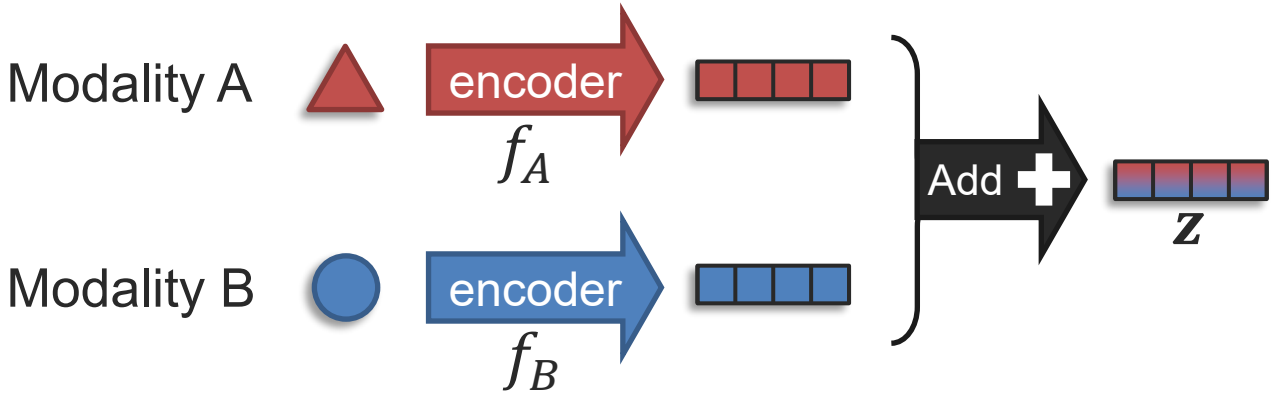


Additive fusion:

$$z = w_1 x_A + w_2 x_B$$

➡ 1-layer neural network can be seen as additive

With unimodal encoders:

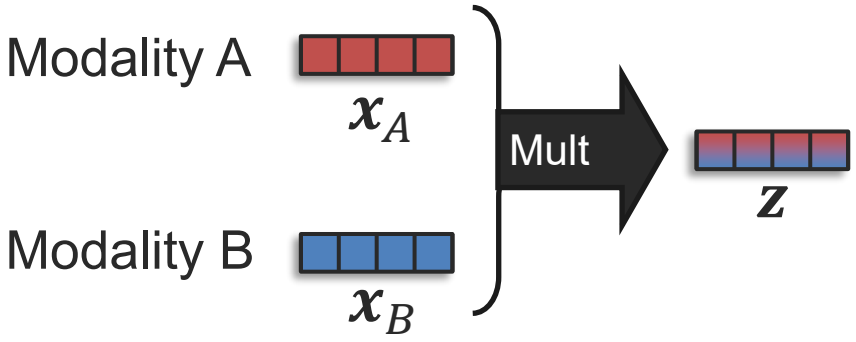


Additive fusion:

$$z = f_A(\blacktriangle) + f_B(\bullet)$$

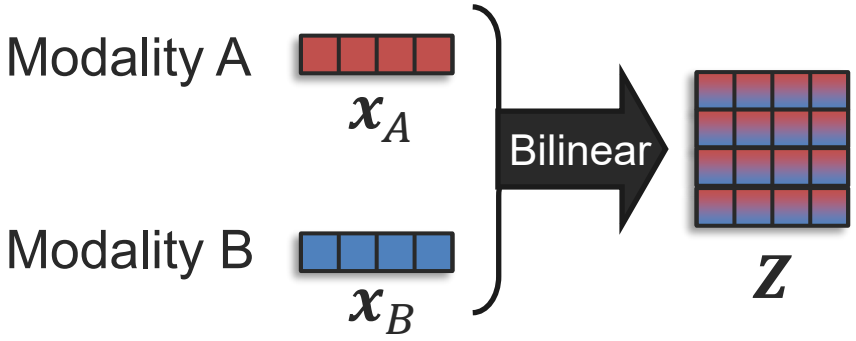
➡ It could be seen as an ensemble approach (late fusion)

Multiplicative Fusion



Simple multiplicative fusion:

$$z = w(x_A \times x_B)$$

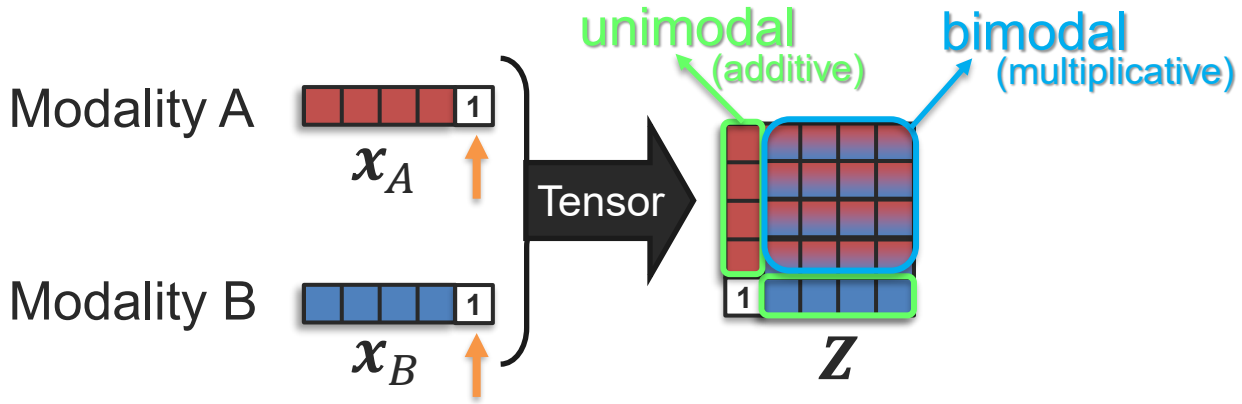


Bilinear Fusion:

$$Z = W(x_A^T \cdot x_B)$$

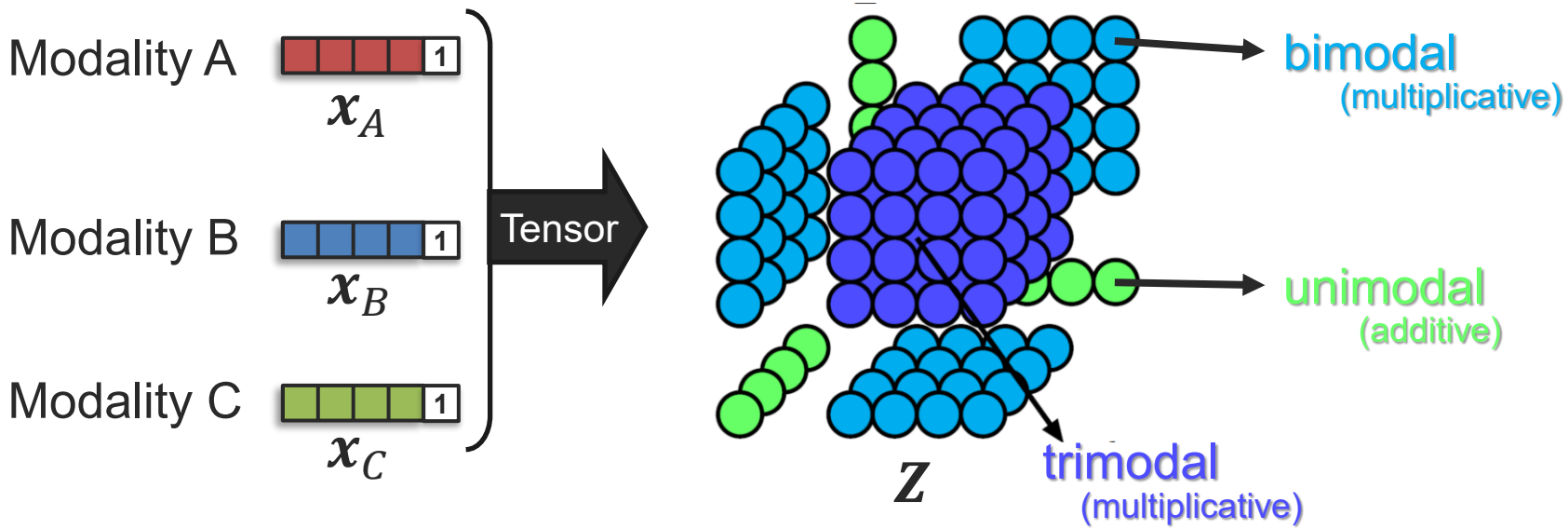
Jayakumar et al., Multiplicative Interactions and Where to Find Them, ICLR 2020

Tensor Fusion



Tensor Fusion (bimodal):

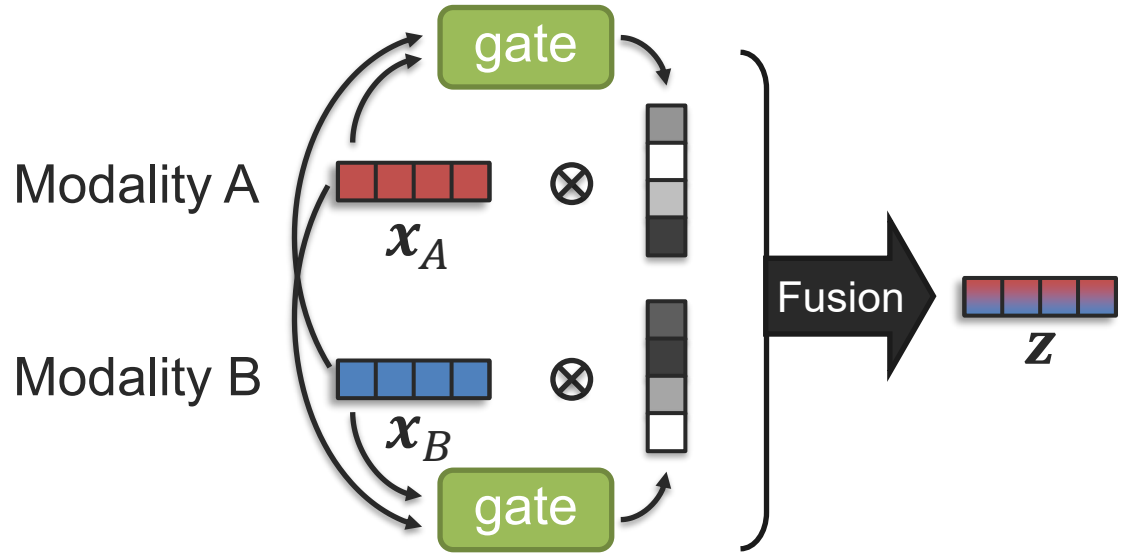
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

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

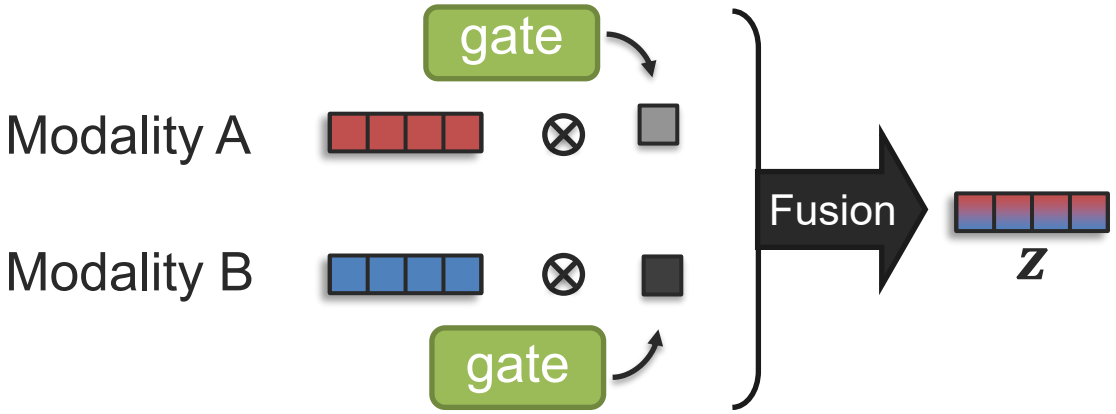
Gated Fusion



Example with additive fusion:

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

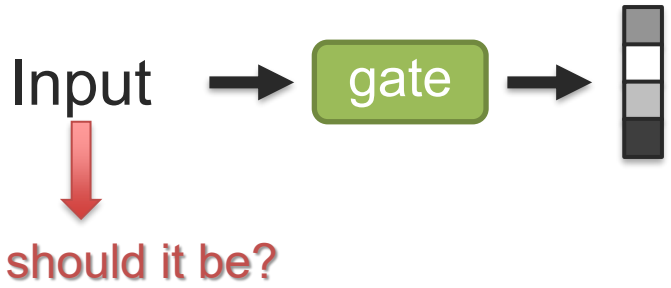
→ g_A and g_B can be seen as attention functions



→ Gating output can be one weight for the whole modality

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

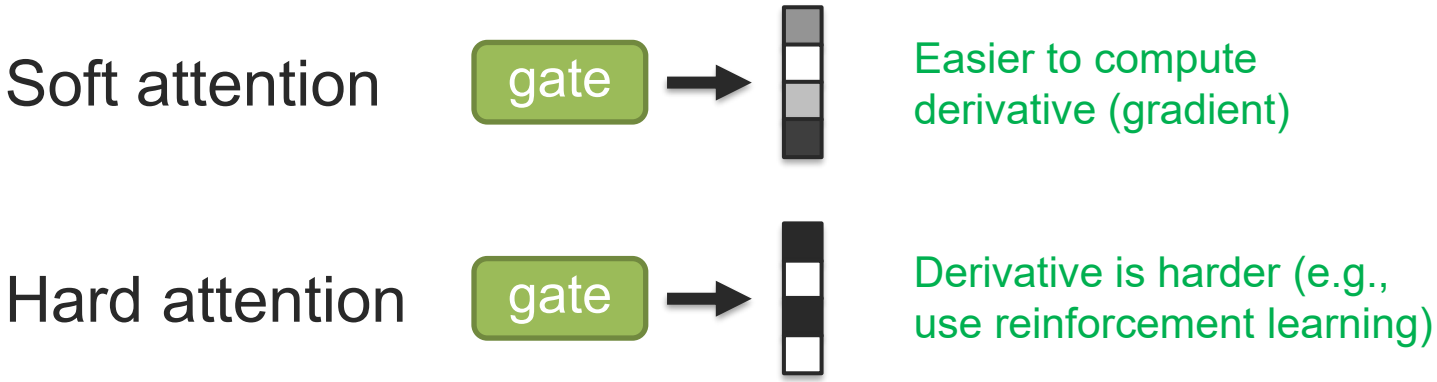
Gating Module (aka, attention module)



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

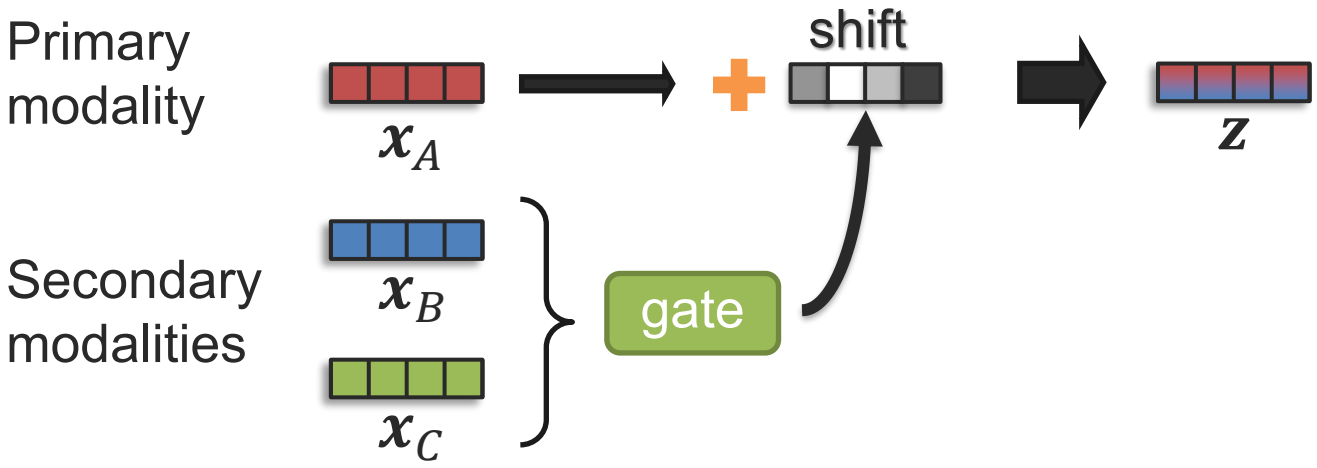
…or with a more positive view:

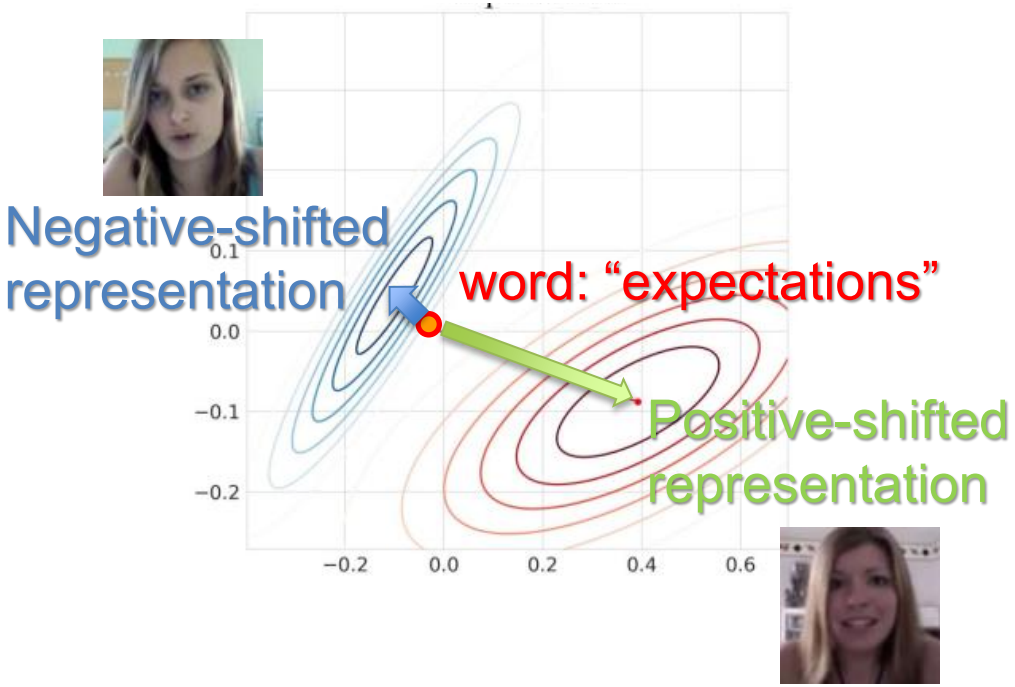
"Neural network designed to select preferable signal to move forward" (attention)



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

Modality-Shifting Fusion





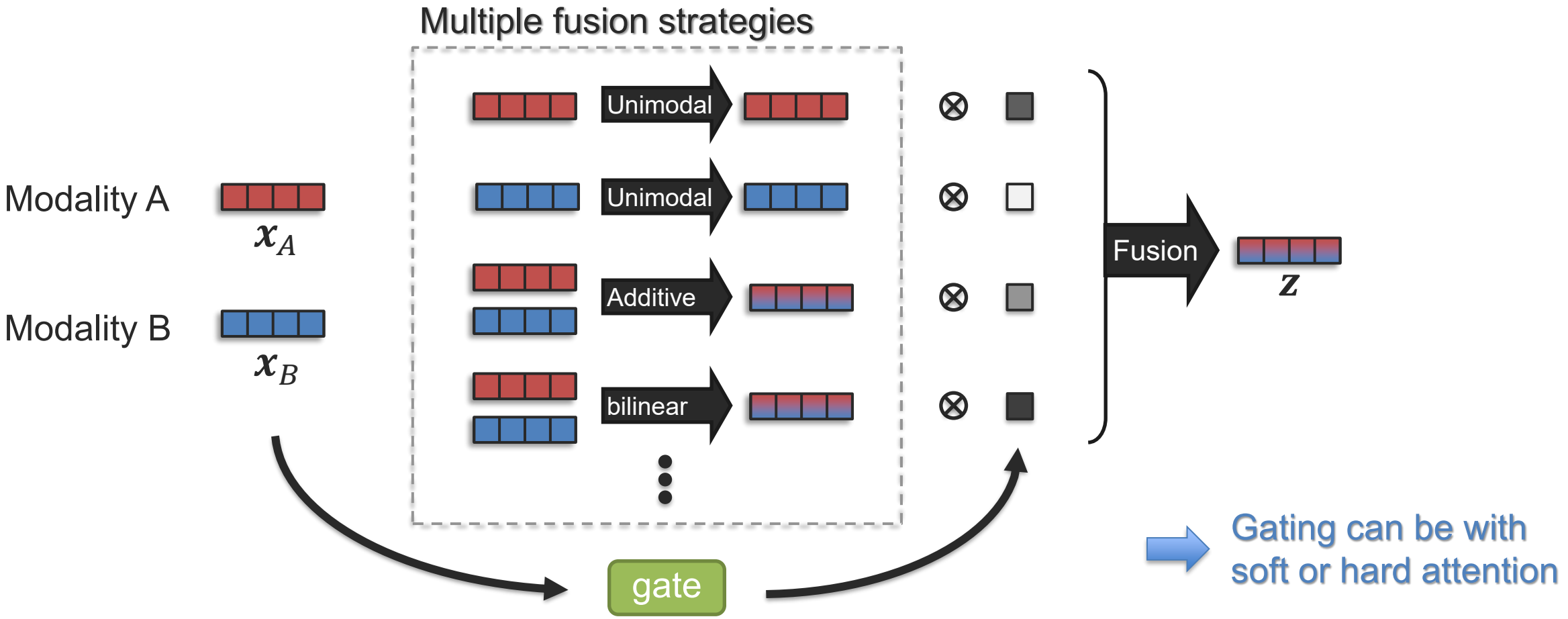
Example with language modality:

Primary modality: language

Secondary modalities: acoustic and visual

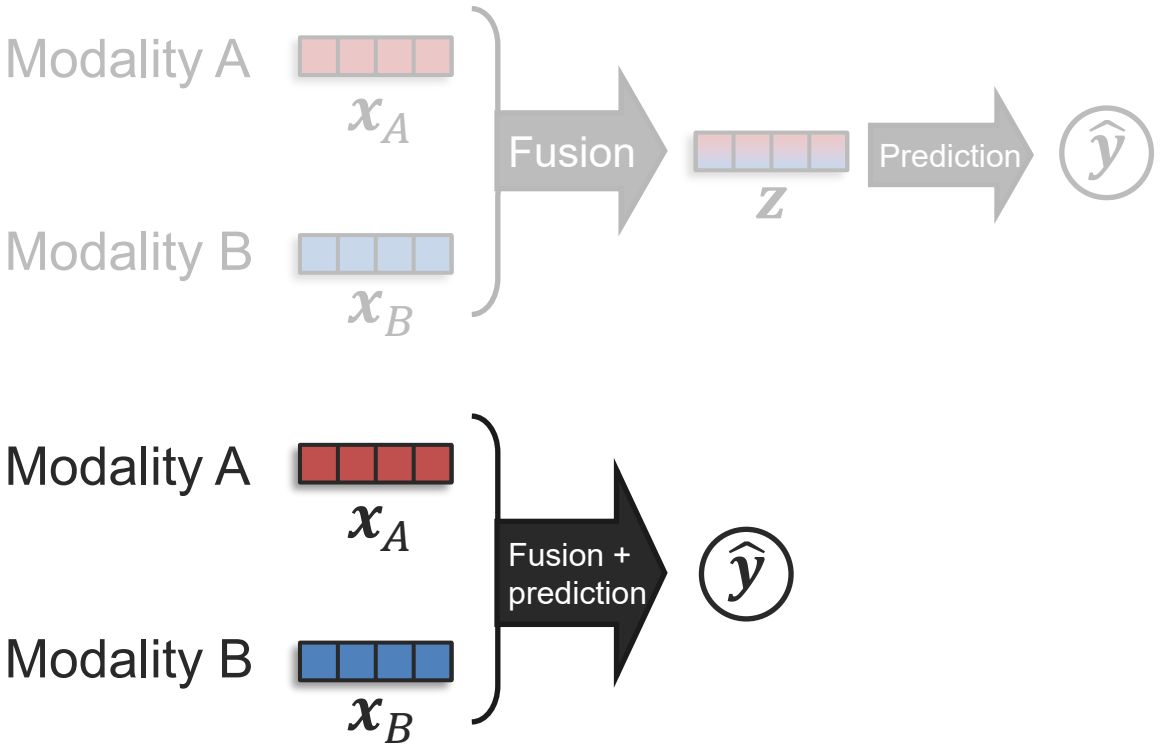
[Wang et al., Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors, AAAI 2019]
[Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers, ACL 2020]

Mixture of Fusions



[Zadeh et al., Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph, ACL 2018]
[Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records, AAAI 2021]

Nonlinear Fusion



Nonlinear fusion:

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}_A, \boldsymbol{x}_B) \in \mathbb{R}^d$$

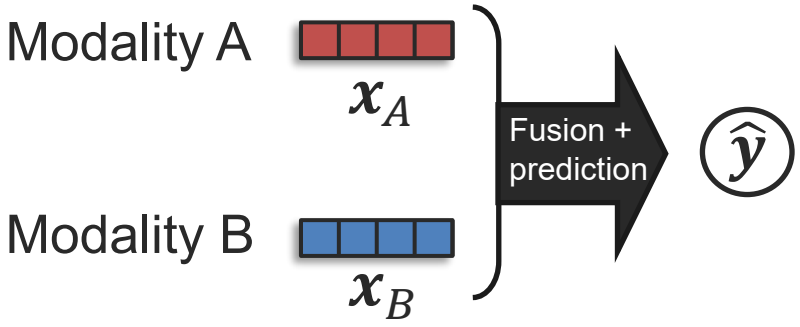
where f could be a multi-layer perceptron or any nonlinear model

➡ This could be seen as *early fusion*:

$$\hat{\boldsymbol{y}} = f([\boldsymbol{x}_A, \boldsymbol{x}_B])$$

… but will our neural network learn the nonlinear interactions?

Measuring Non-Additive Interactions



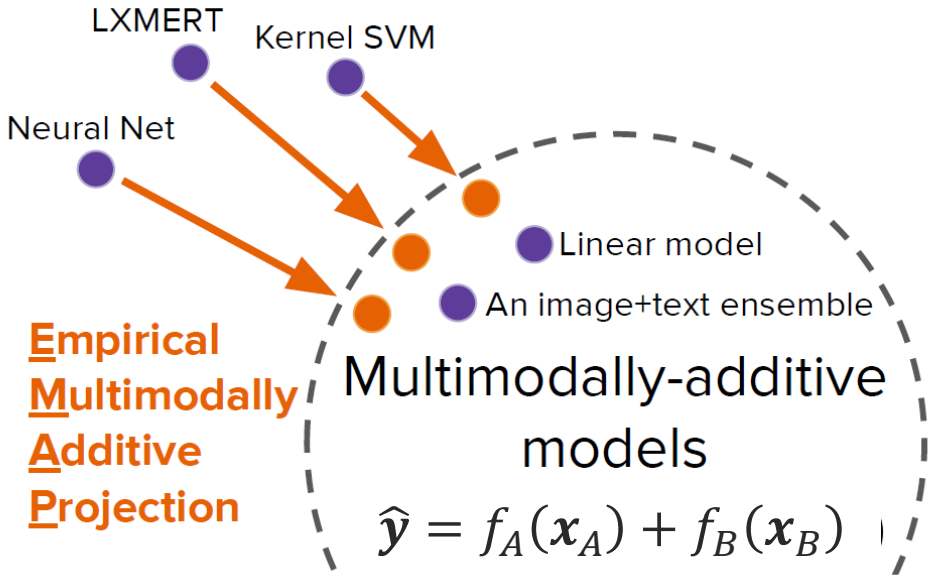
Nonlinear fusion:

$$\hat{y} = f(x_A, x_B)$$

Projection?

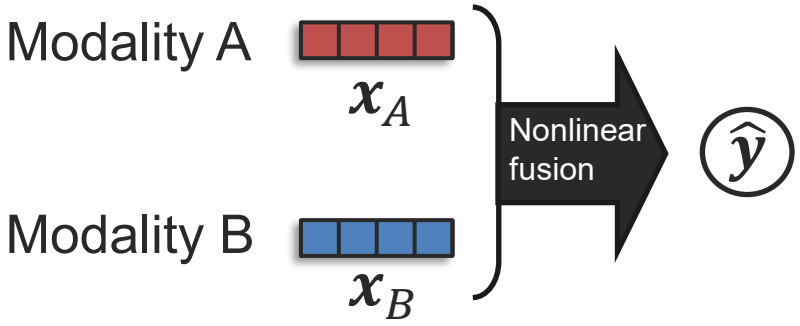
Additive fusion:

$$\hat{y} = f_A(x_A) + f_B(x_B)$$



Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, EMNLP 2020 → introduced the EMAP method

Measuring Non-Additive Interactions



Nonlinear fusion:

$$\hat{y} = f(x_A, x_B)$$

Projection?

Additive fusion:

$$\hat{y}' = f_A(x_A) + f_B(x_B)$$

Projection from nonlinear to additive (using EMAP):

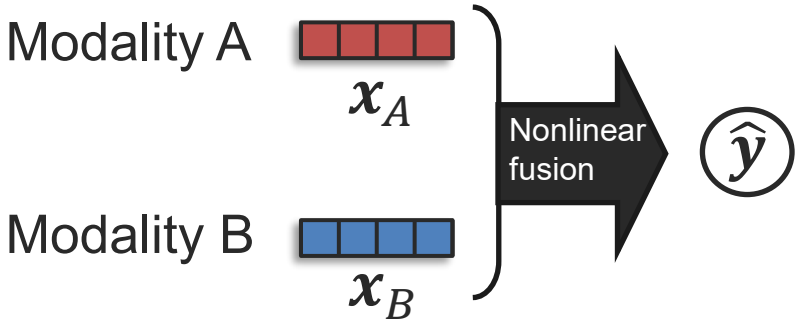
$$\tilde{f}(x_A, x_B) = \underbrace{\mathbb{E}_{x_B}[f(x_A, x_B)]}_{f_A(x_A)} + \underbrace{\mathbb{E}_{x_A}[f(x_A, x_B)]}_{f_B(x_B)}$$

Modality A + Modality B

Additive fusion (approximation)

[Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, EMNLP 2020]

Measuring Non-Additive Interactions



Nonlinear fusion:

$$\hat{y} = f(x_A, x_B)$$

EMAP projection

Additive fusion:

$$\hat{y}' = \hat{f}_A(x_A) + \hat{f}_B(x_B) + \hat{\mu}$$

		I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Nonlinear	Neural Network	90.4	69.2	78.5	51.1	63.5	71.1	79.9
Polynomial	Polykernel SVM	**91.3**	**74.4**	**81.5**	50.8	–	72.1	**80.9**
Nonlinear	FT LXMERT	83.0	68.5	76.3	**53.0**	63.0	66.4	78.6
Nonlinear	↳ + Linear Logits	89.9	73.0	80.7	**53.4**	**64.1**	**75.5**	80.3
Additive	Linear Model	90.4	72.8	80.9	51.3	63.7	**75.6**	76.1
	Best Model	**91.3**	**74.4**	**81.5**	**53.4**	***64.2***	**75.5**	**80.9**
Additive	↳ + EMAP	**91.1**	**74.2**	**81.3**	51.0	***64.1***	**75.9**	**80.7**

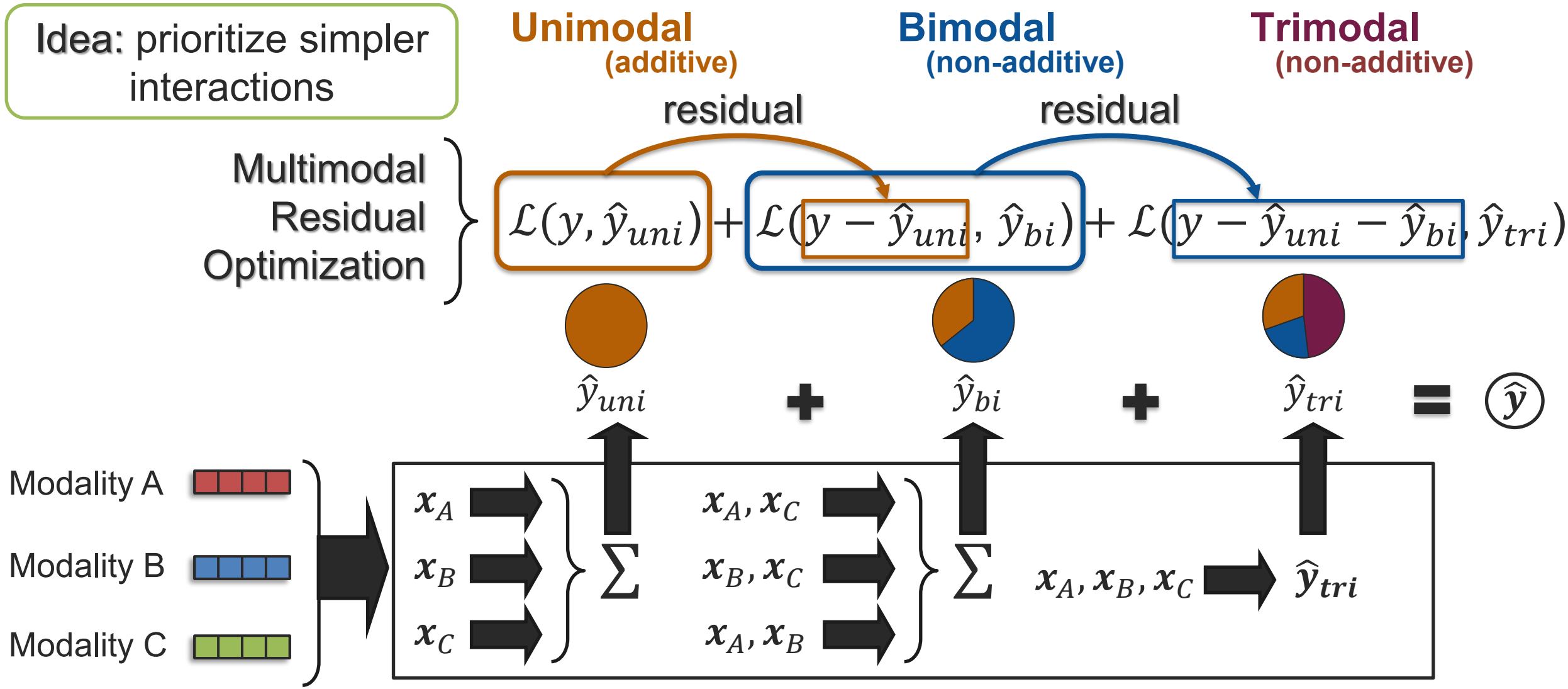
Always a good baseline!

Differences are small!!!

[Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, EMNLP 2020]

Learning Non-additive Bimodal and Trimodal Interactions

Idea: prioritize simpler interactions



[Wortwein et al., Beyond Additive Fusion: Learning Non-Additive Multimodal Interactions, Findings-EMNLP 2022]

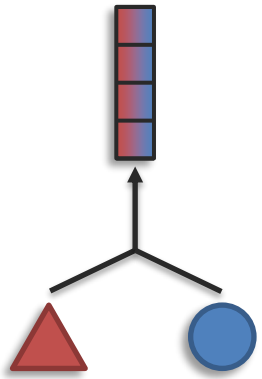
Representation Coordination

Challenge 1: Representation

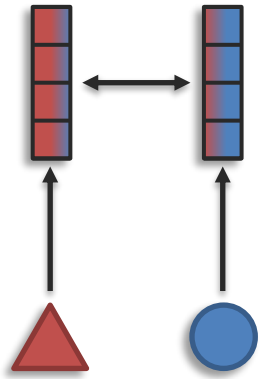
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

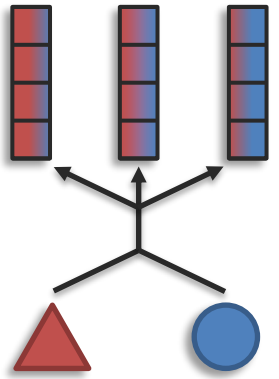
Fusion



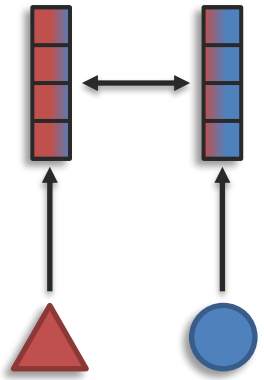
Coordination



Fission

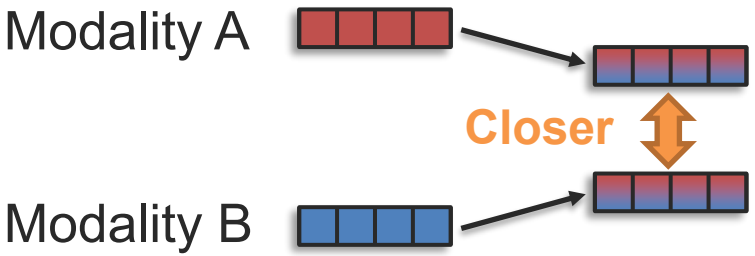


Sub-Challenge 1b: Representation Coordination

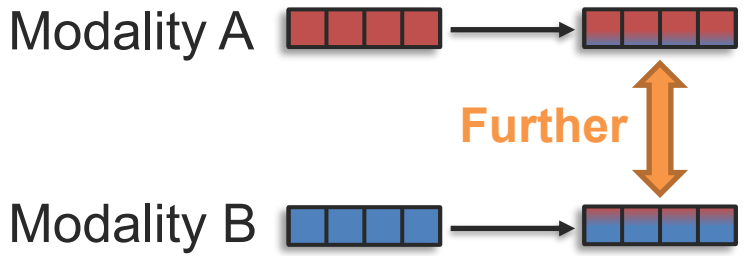


Definition: Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions

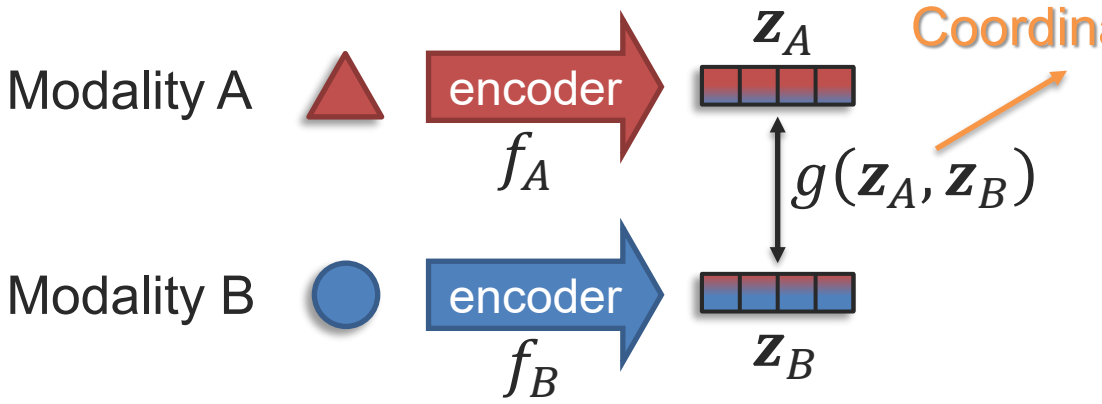
Strong Coordination:



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:

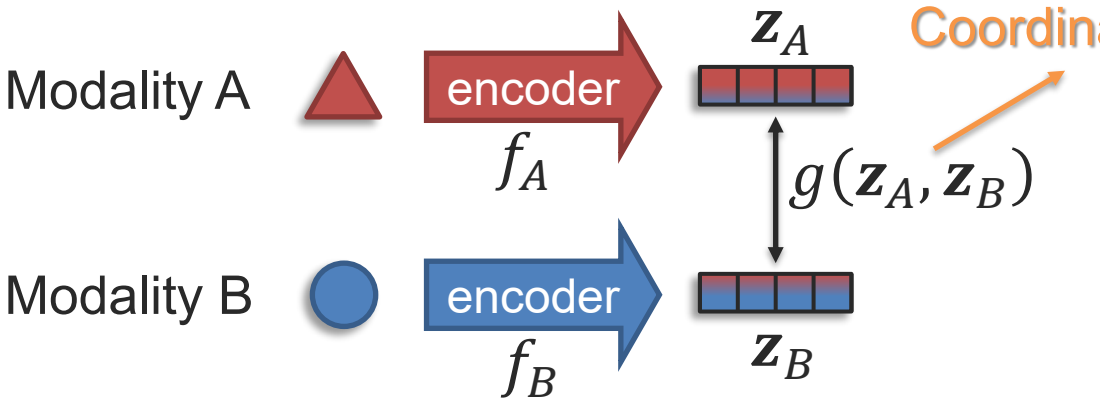
① Cosine similarity:

$$g(z_A, z_B) = \frac{z_A \cdot z_B}{\|z_A\| \|z_B\|}$$

Strong coordination!

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

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}

Examples of coordination function:

② Kernel similarity functions:

$$g(z_A, z_B) = k(z_A, z_B)$$

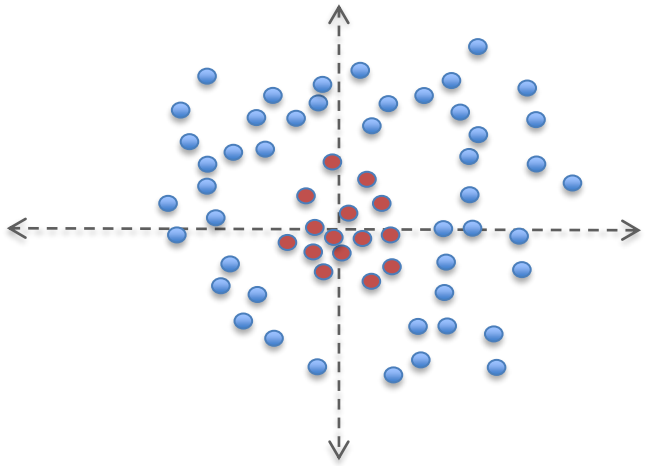
- Linear
- Polynomial
- Exponential
- RBF

➡ All these examples bring relatively strong coordination between modalities

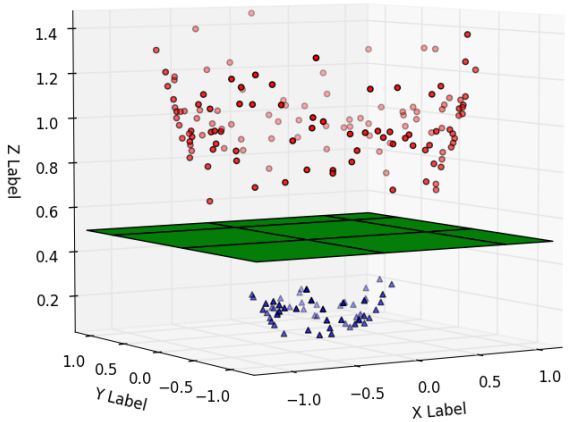
Kernel Function

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!



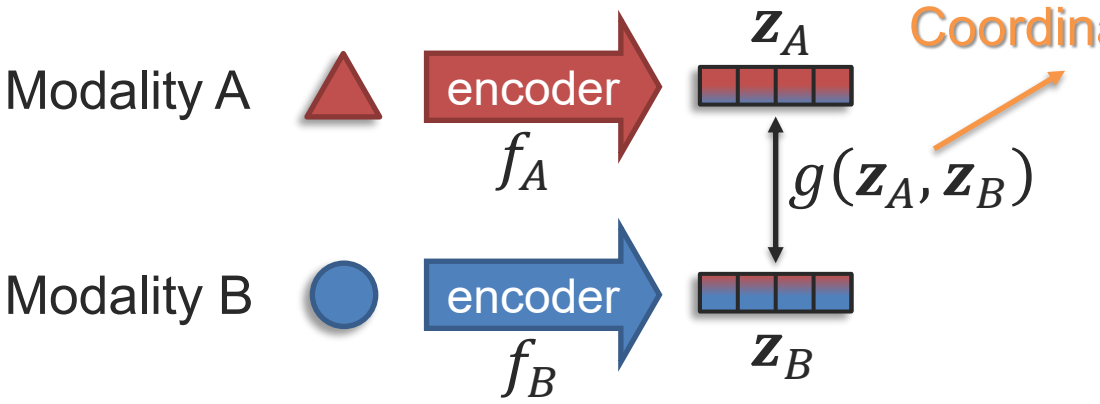
Not linearly separable in x space



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:

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

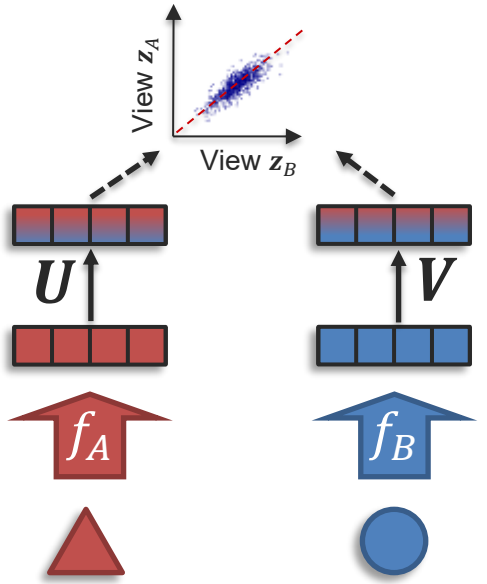
with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

③ Canonical Correlation Analysis (CCA):

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

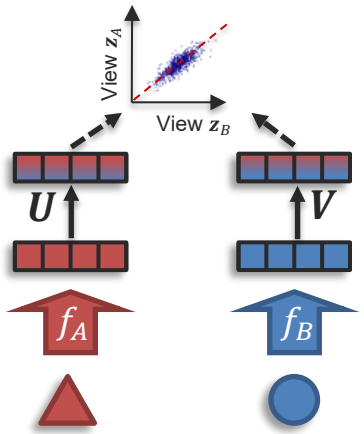
➡ CCA includes multiple projections, all orthogonal with each others

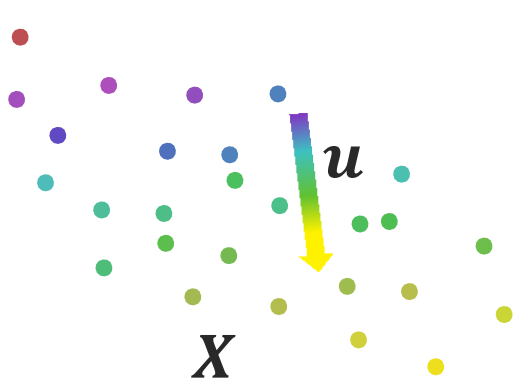


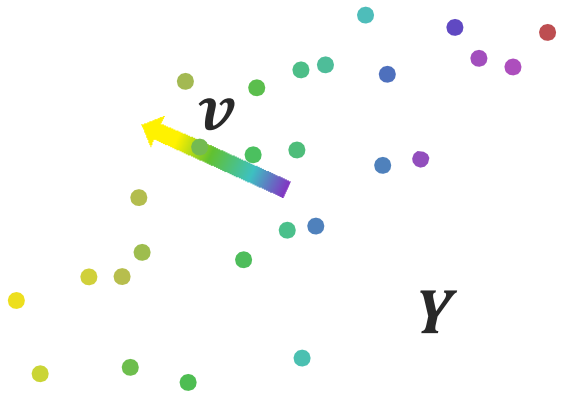
Correlated Projection

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

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



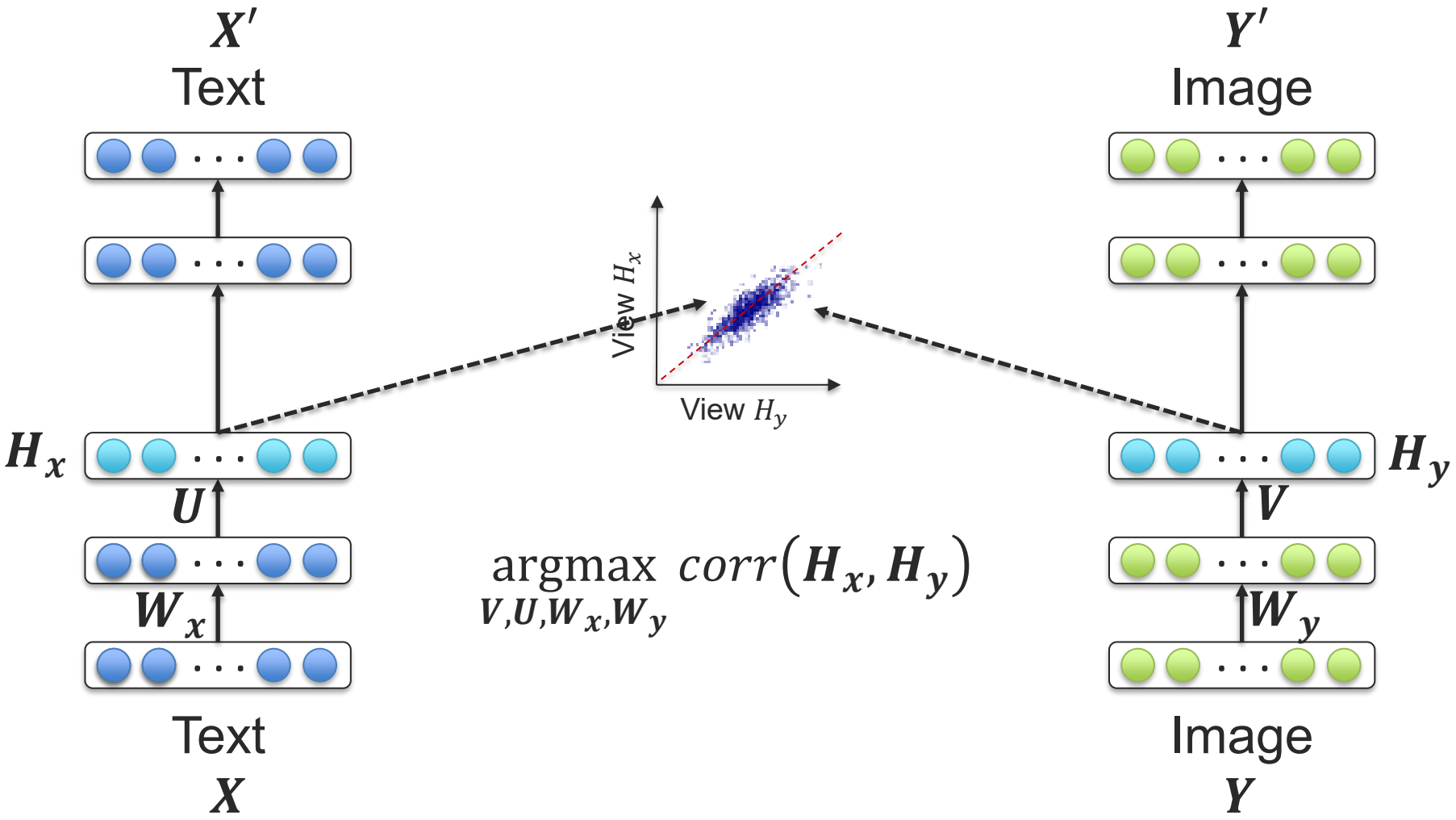




Two views $\boldsymbol{X}, \boldsymbol{Y}$ where same instances have the same color

➡ Remember that X and Y consist of paired data

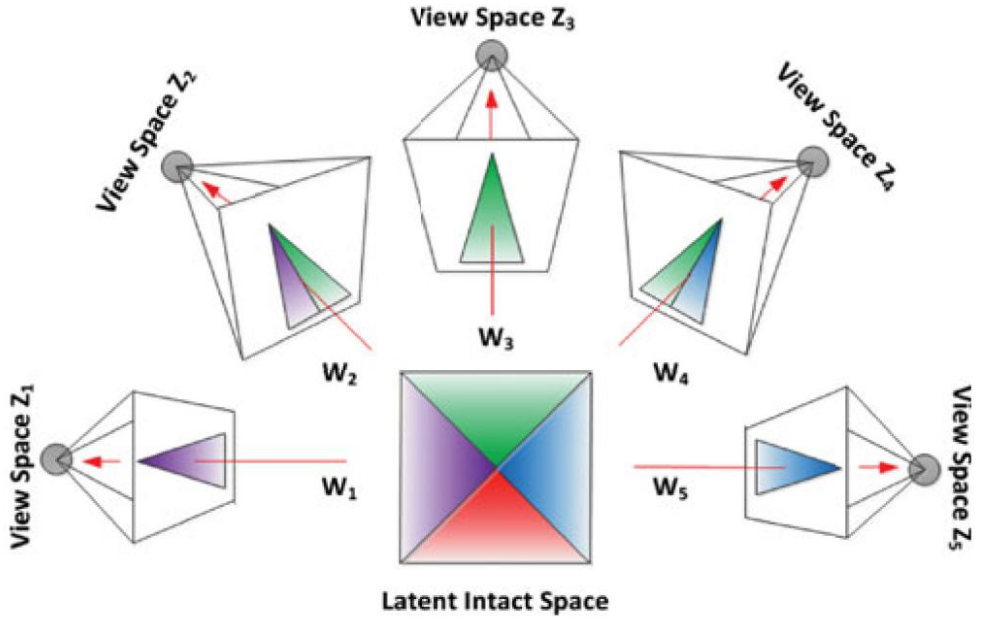
Deep Canonically Correlated Autoencoders (DCCAE)



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

Multi-view Latent "Intact" Space

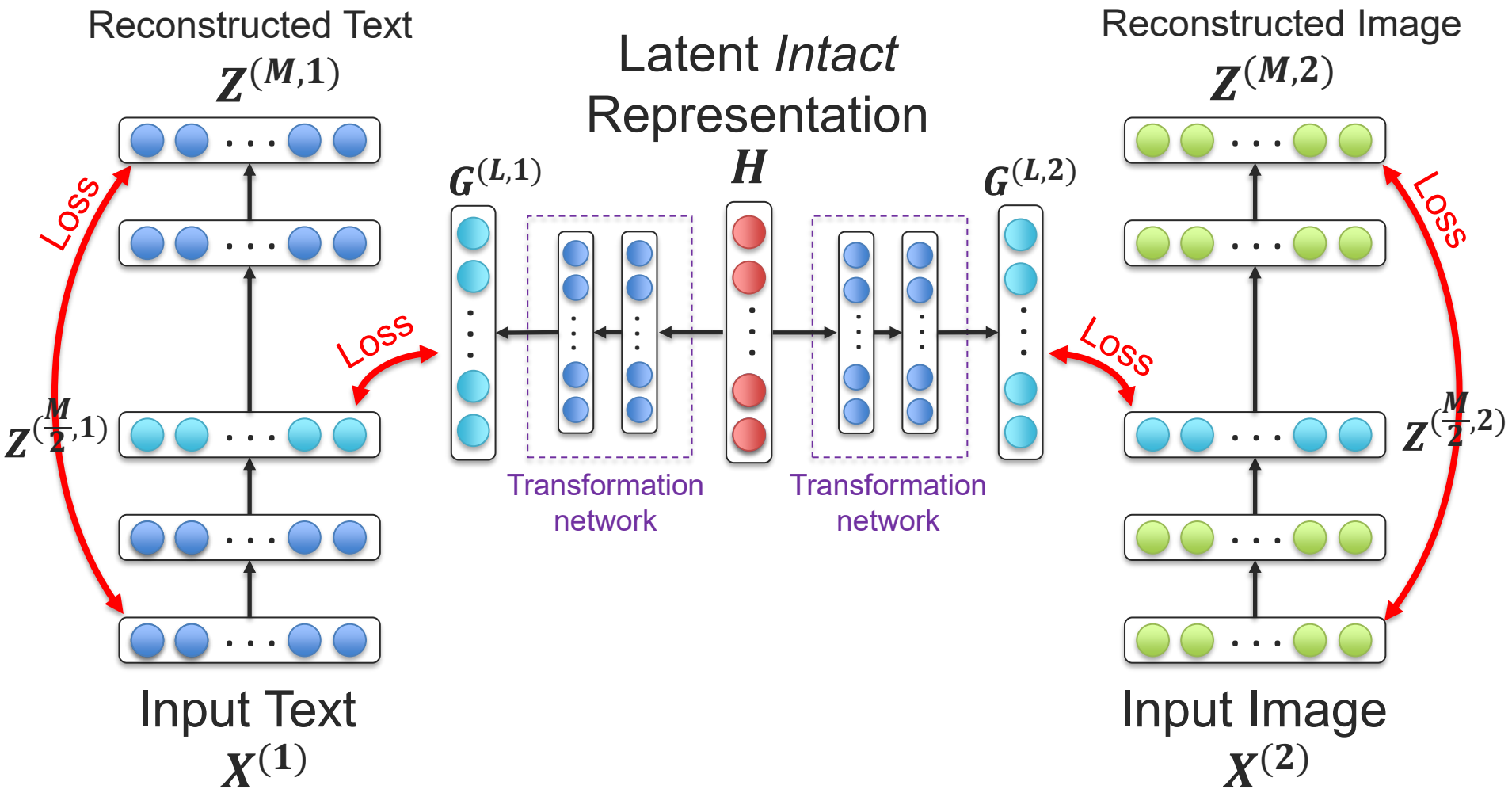
Given multiple views z_i from the same "object":



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

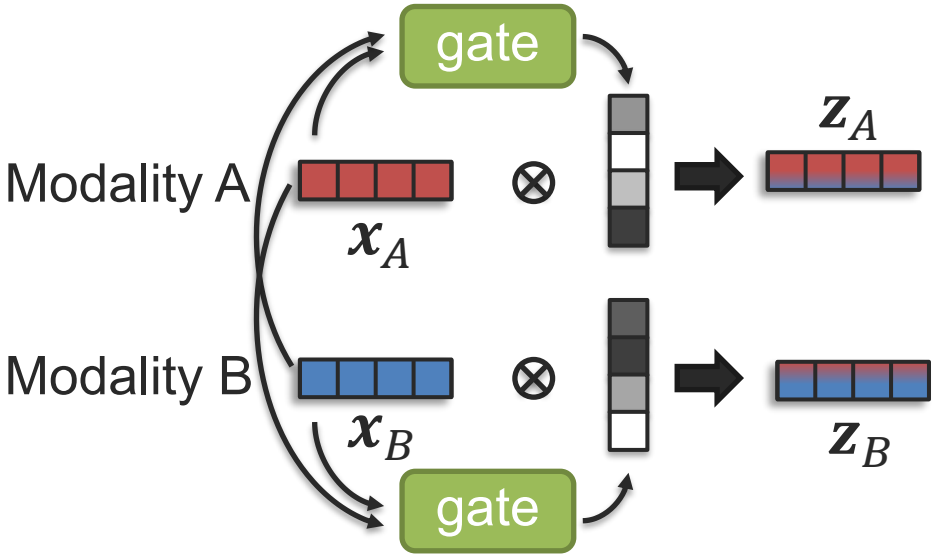
Xu et al., Multi-View Intact Space Learning, TPAMI 2015

Auto-Encoder in Auto-Encoder Network



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

Gated Coordination



Gated coordination:

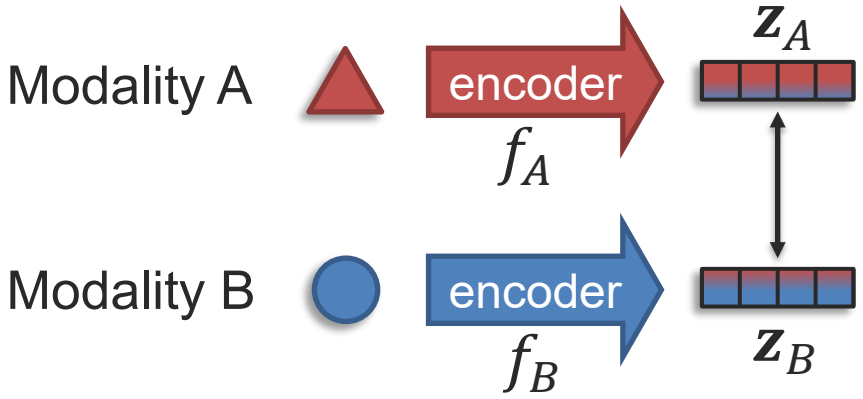
$$z_A = g_A(x_A, x_B) \cdot x_A$$

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

Related to attention modules in transformers

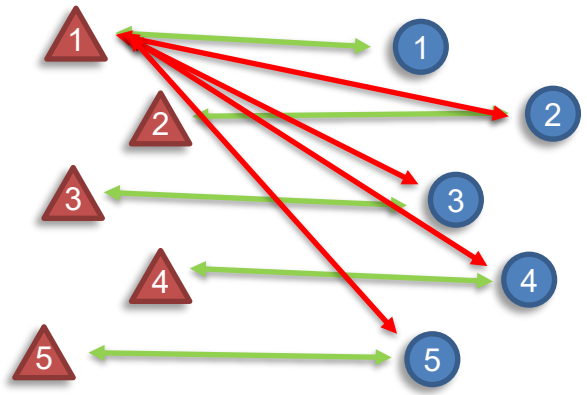
More about it next week!

Coordination with Contrastive Learning



Paired data: { ▲ , ● }

(e.g., images and text descriptions)



Positive pairs

Negative pairs

Contrastive loss:

➡ brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

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

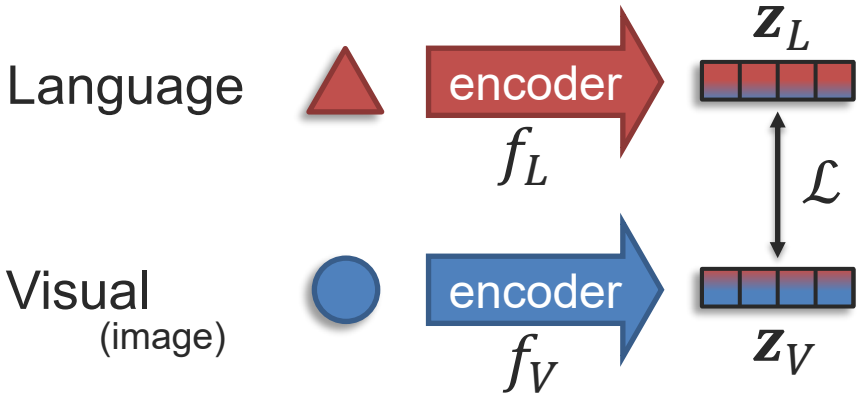
positive pairs

negative pair

Similarity functions are often cosine similarity

➡ Similar to hinge loss

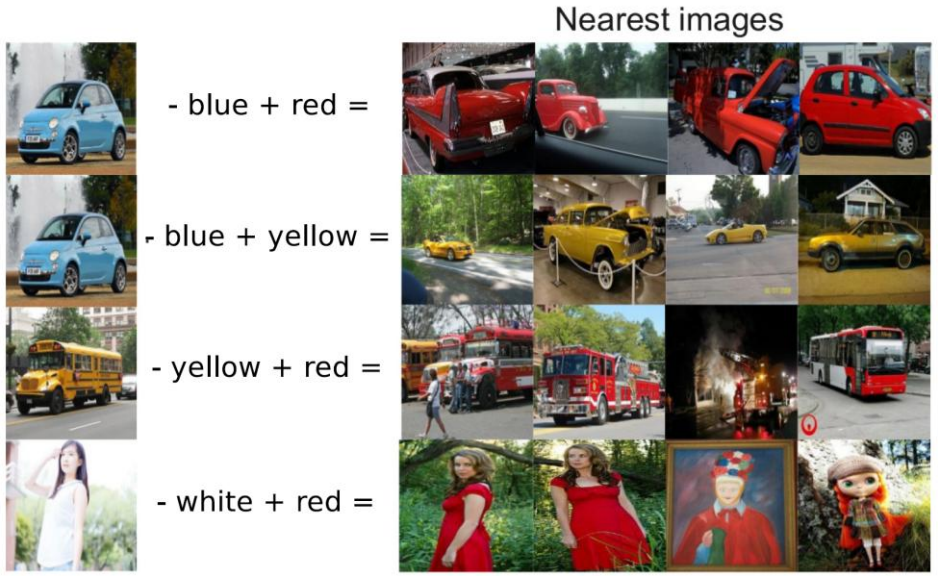
Example – Visual-Semantic Embeddings

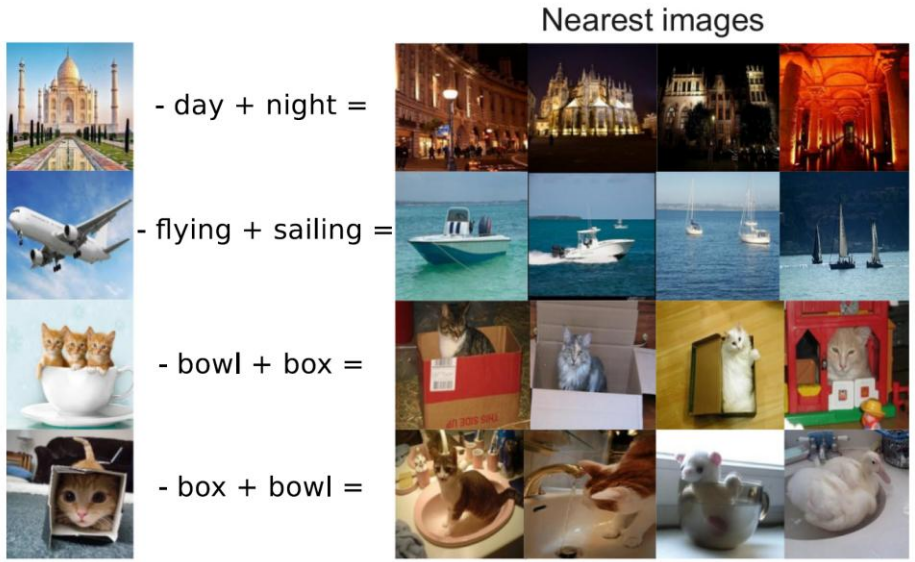


Two contrastive loss terms:

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

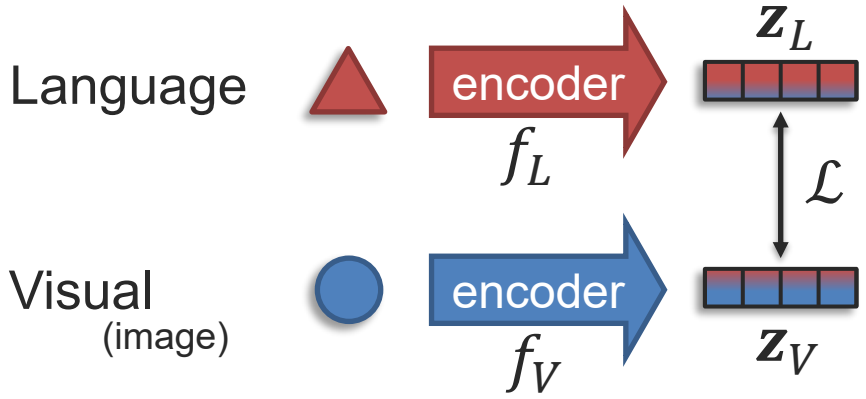
$$+ \max\{0, \alpha + sim(\boldsymbol{z}_V, \boldsymbol{z}_L^+) - sim(\boldsymbol{z}_V, \boldsymbol{z}_L^-)\}$$





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

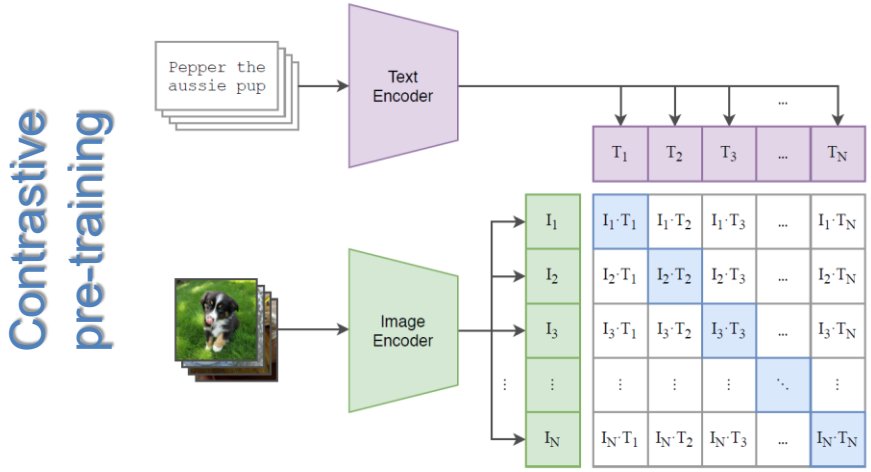
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\text{sim}(\boldsymbol{z}_A^l, \boldsymbol{z}_B^l)}{\sum_{i=1}^{N} \text{sim}(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

positive pairs

negative pairs and positive pairs

Similarity function can be cosine similarity

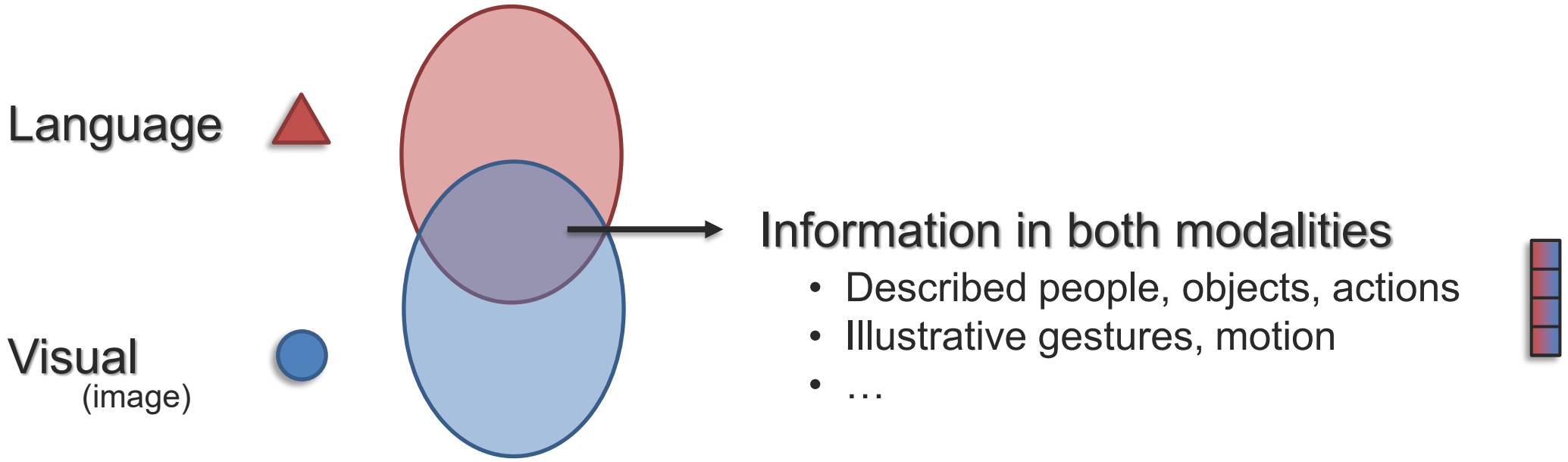
Positive and negative pairs:



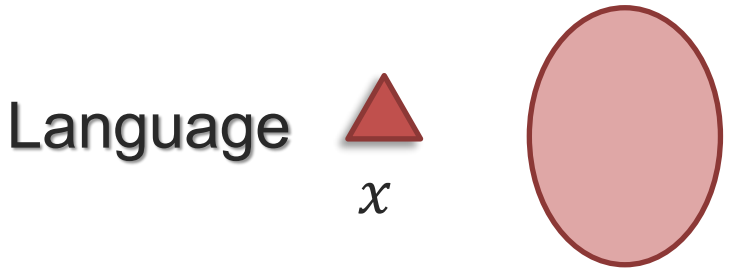
- CLIP encoders (f_L and f_V) are great for language-vision tasks
- $\boldsymbol{z}_L$ and $\boldsymbol{z}_V$ are coordinated but not identical representation spaces

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

Multimodal Coordination – Information Theory



Information and Entropy – Information Theory



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 $I(x)$

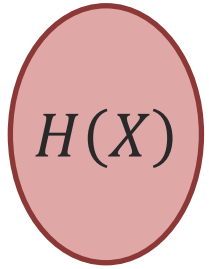
$$I(x) \sim \frac{1}{p(x)}$$

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

Shannon, A Mathematical Theory of Communication, 1948

Information and Entropy – Information Theory

Language x



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

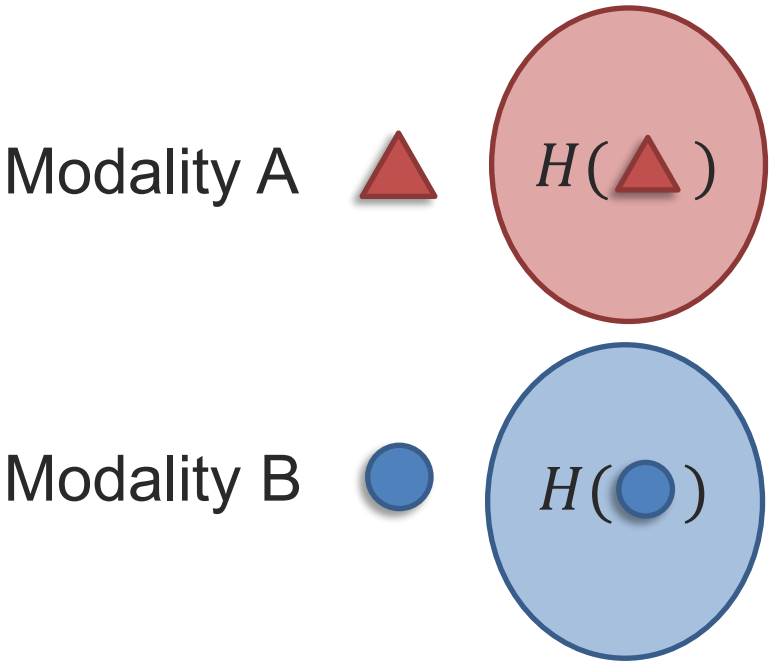
Entropy: weighted average of all possible outcomes from $\mathcal{X}$

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

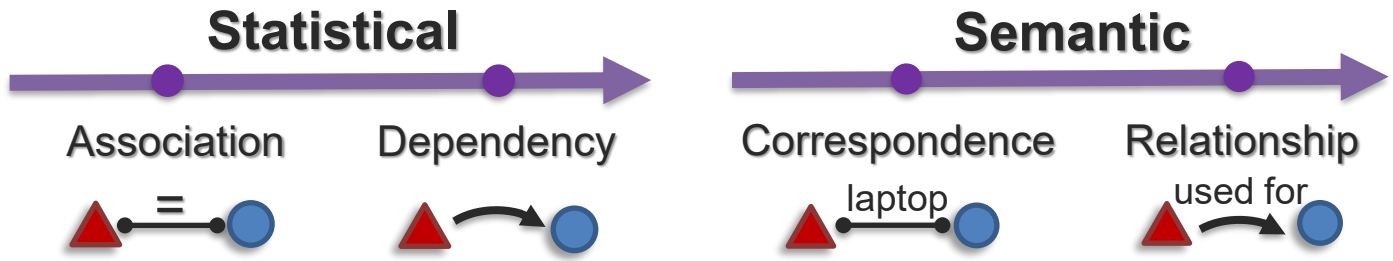
➡ Entropy can also be defined for continuous random variables

Entropy with Two Modalities

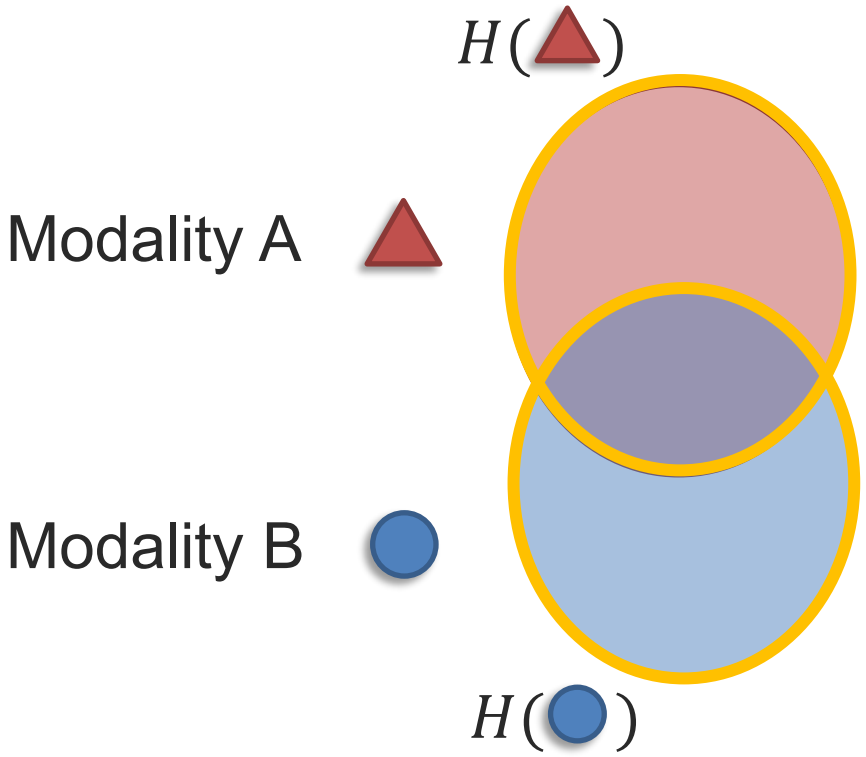
If no overlapping information



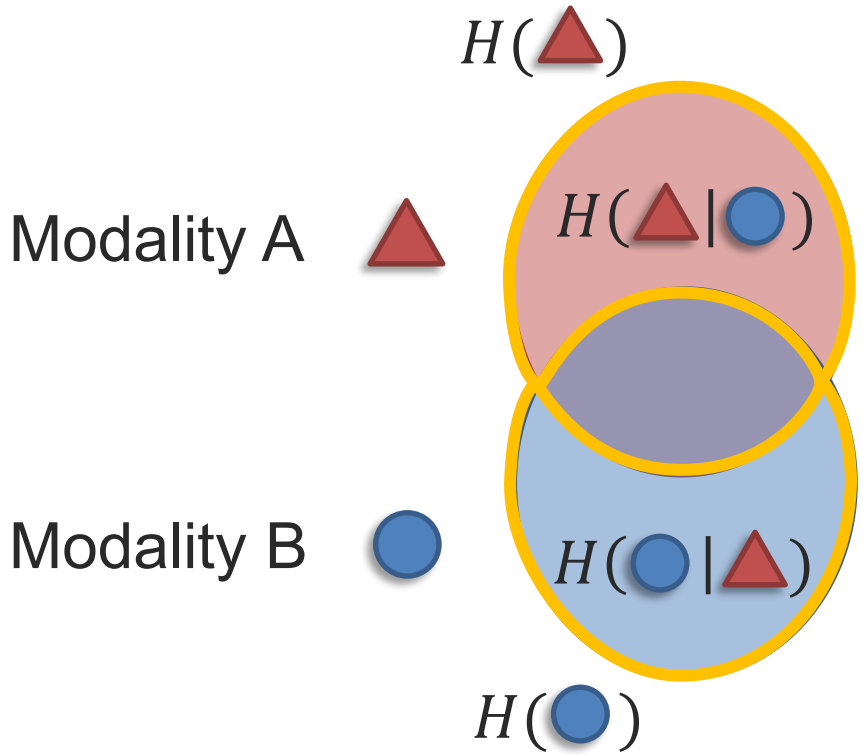
But in most real-world scenarios, modalities are *inter-connected*



Entropy with Two Modalities



Entropy with Two Modalities



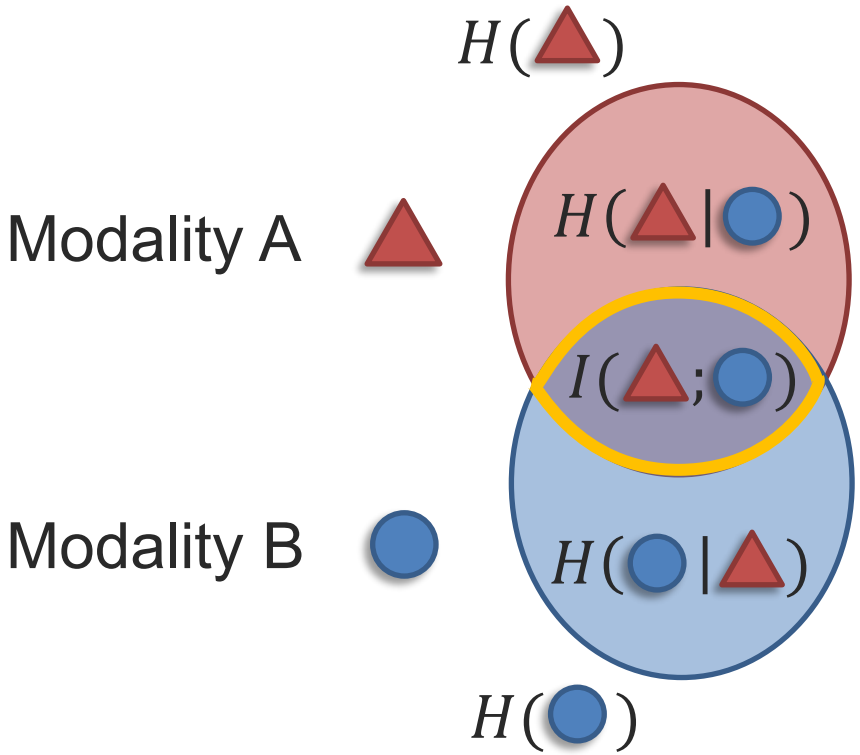
Conditional entropy $H(Y|X)$

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

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

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

Entropy with Two Modalities



Mutual information $I(X;Y)$

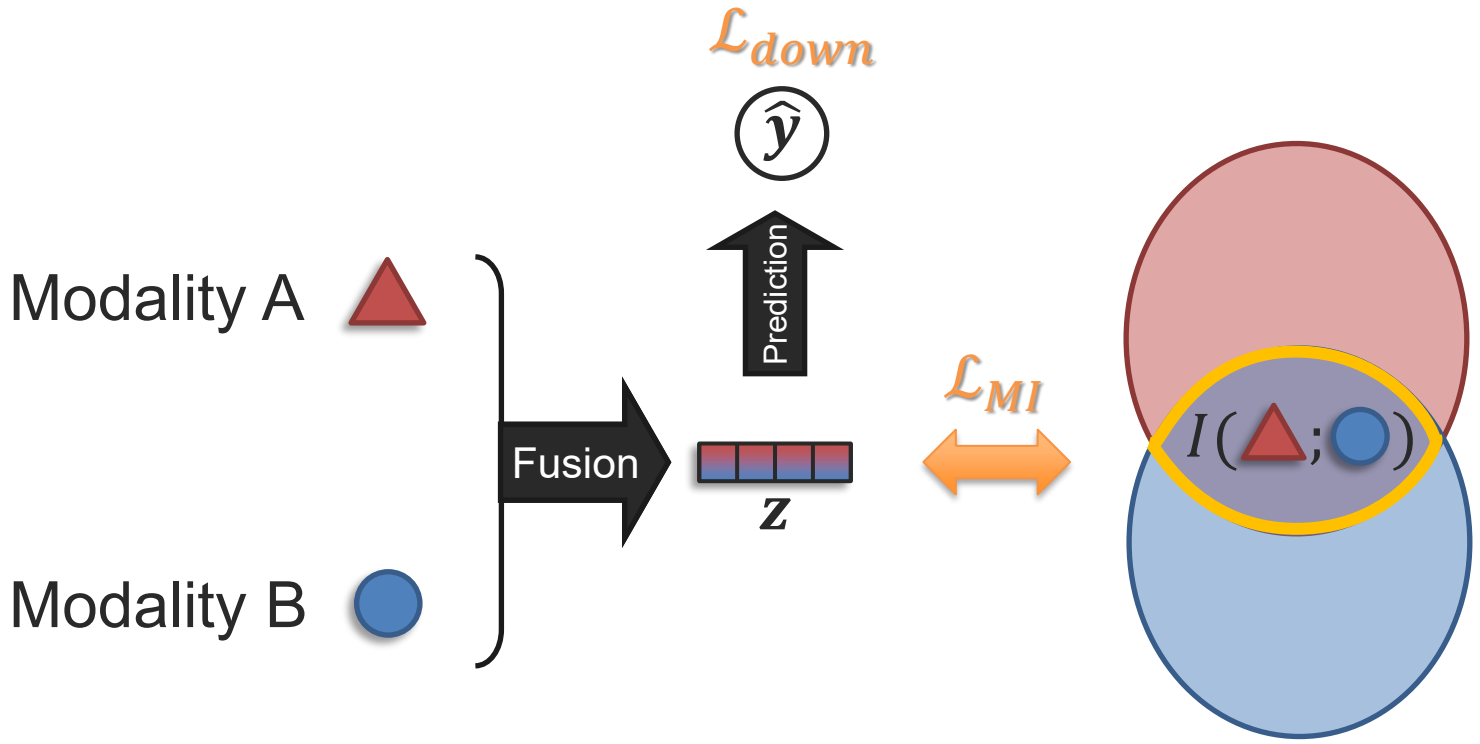
$$I(X;Y) = H(X) - H(X|Y)$$

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

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

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

Multimodal Fusion with Mutual Information

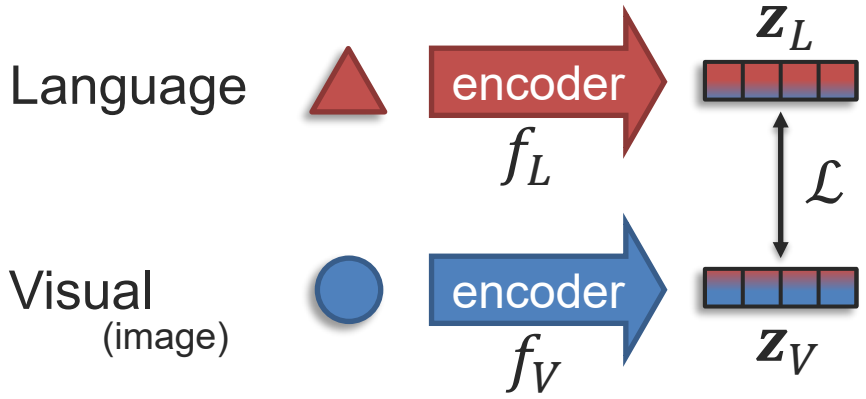


Assumption?

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

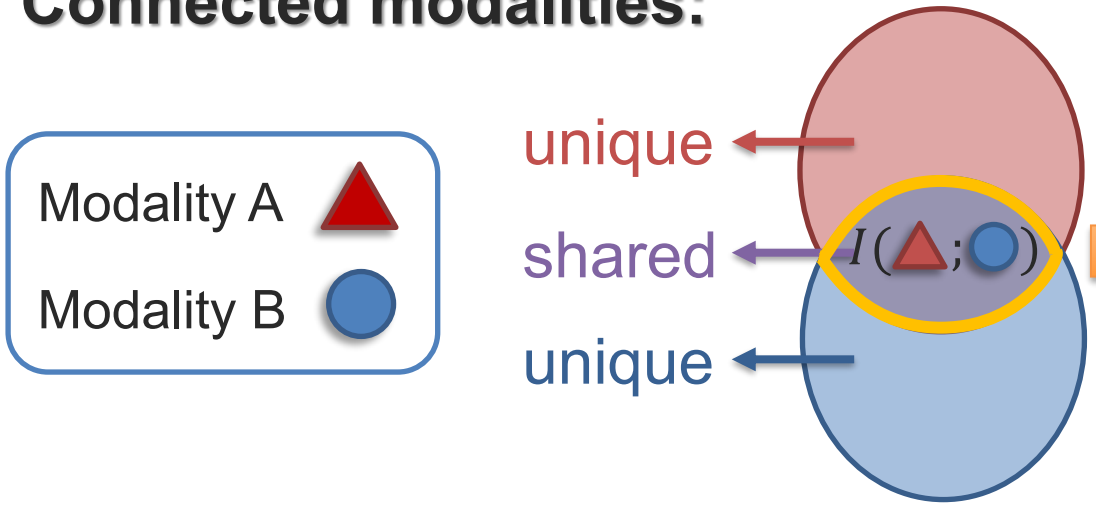


Popular contrastive loss: InfoNCE

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

CLIP focuses on shared connections

Connected modalities:



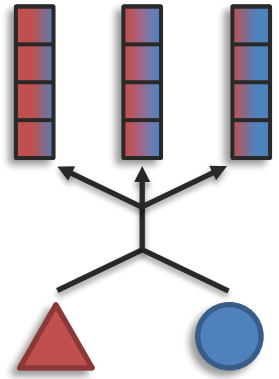
Mutual information $I(X;Y)$

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

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

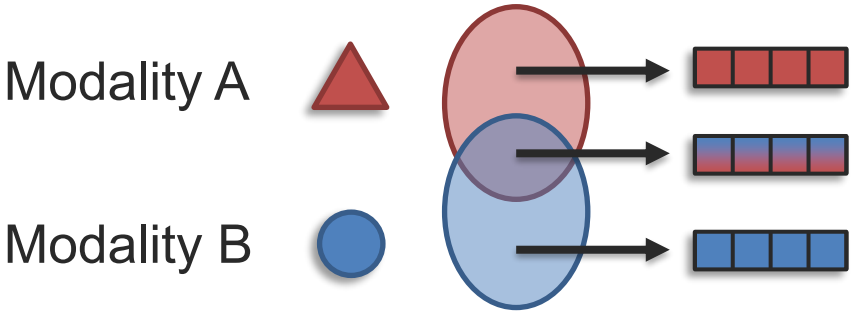
Representation Fission

Sub-Challenge 1c: Representation Fission

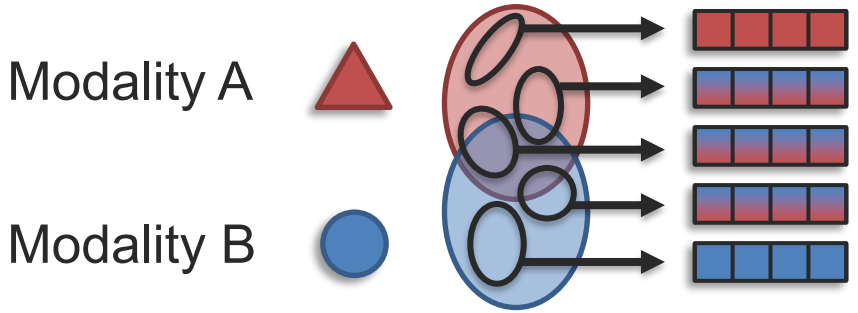


Definition: Learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

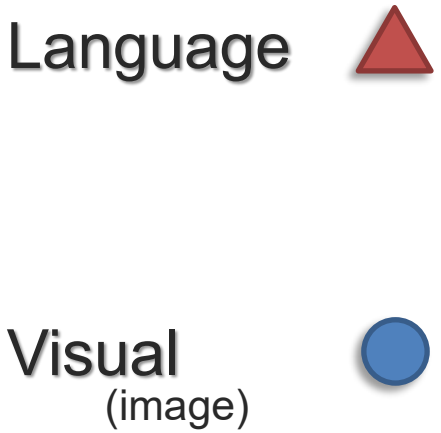
Modality-level fission:

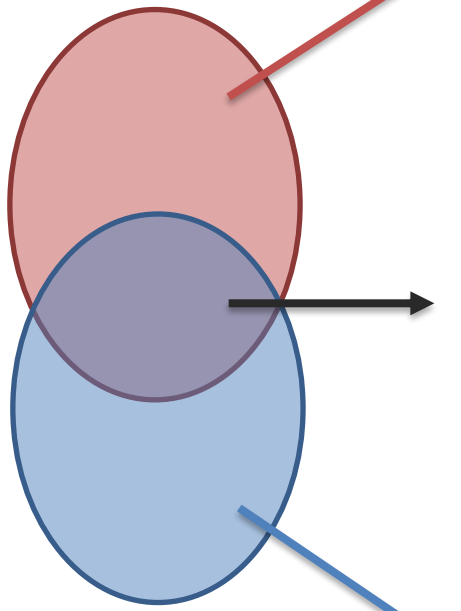


Fine-grained fission:



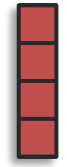
Modality-Level Fission





Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …



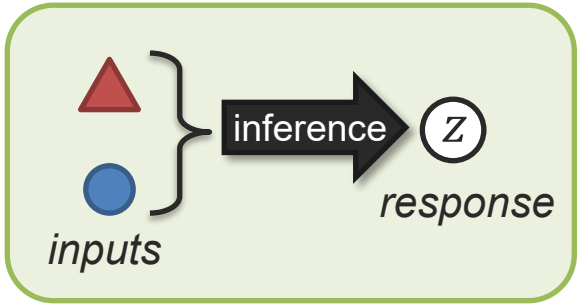
Information in both modalities

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

Information primarily in visual modality

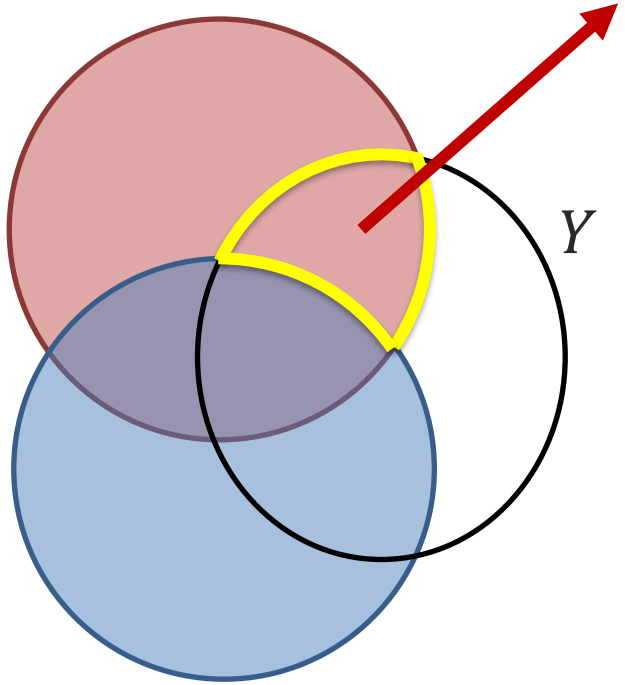
- Texture, visual appearance
- Depth, perspective, motion
- …

Modeling Multimodal Interactions



Modality A

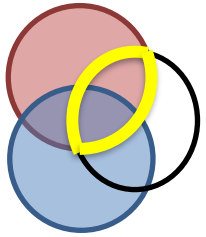
Modality B



Task-relevant **unique** information

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

Mutual information…



$I(\blacktriangle; Y)$

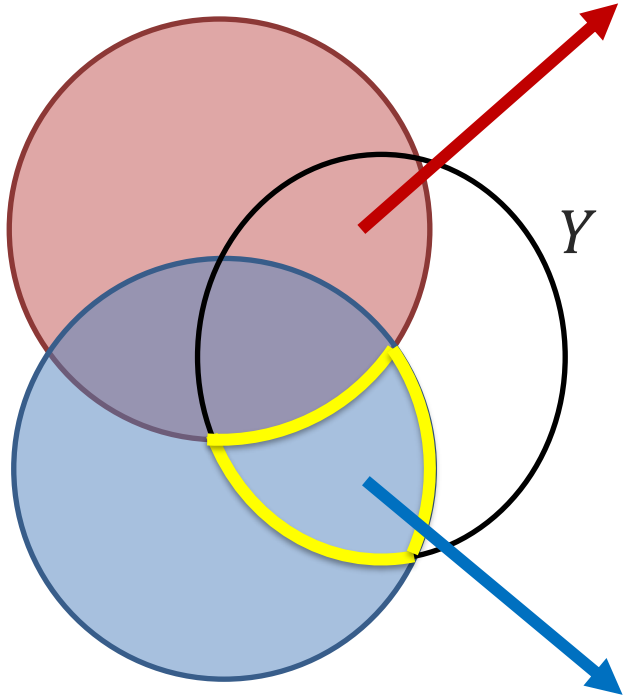
…but without

Information theory as a framework for modeling multimodal interactions between input modalities and responses

Modeling Multimodal Interactions

Modality A

Modality B



Task-relevant **unique** information

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

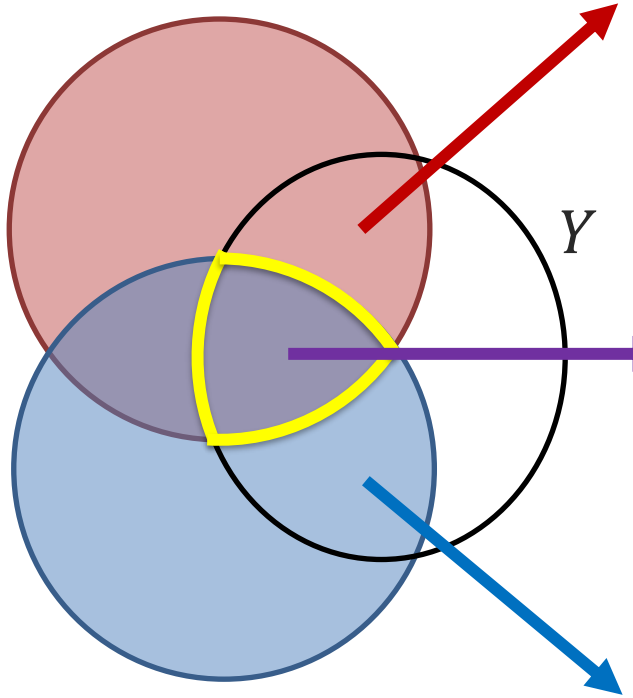
Task-relevant **unique** information

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

Modeling Multimodal Interactions

Modality A

Modality B



Task-relevant **unique** information

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

Task-relevant **shared** information

$I(\blacktriangle; \bullet; Y)$

…but how to compute this 3-way mutual information?

Task-relevant **unique** information

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

Modeling Multimodal Interactions

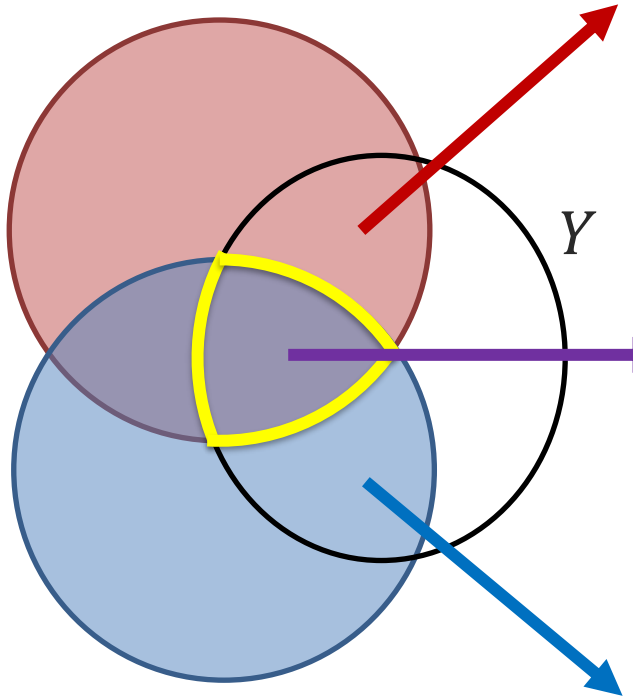
Partial Information Decomposition (PID)

$I(\triangle;\bigcirc;Y) = R - S$

factorizes 3-way mutual information into:

R: redundancy

S: Synergy



Task-relevant **unique** information

$I(\triangle;Y|\bigcirc)$

Task-relevant **shared** information

$I(\triangle;\bigcirc;Y)$

…but how to compute 3-way mutual information?

Task-relevant **unique** information

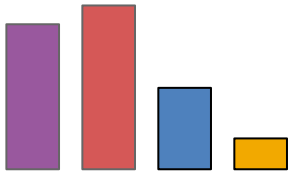
$I(\bigcirc;Y|\triangle)$

More details:
Liang et al., Quantifying & Modeling Feature Interactions: An Information Decomposition Framework. Neurips 2023

Quantifying Multimodal Interactions

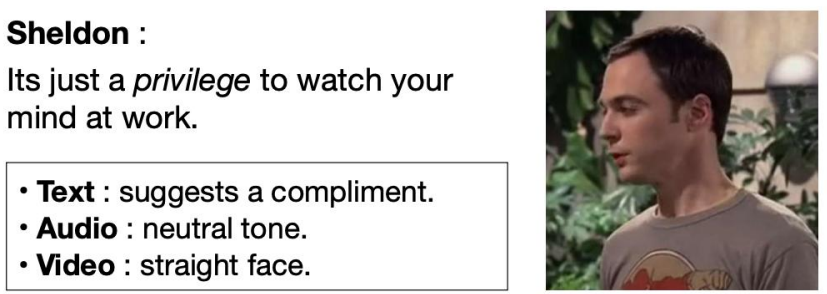
These interactions can be efficiently estimated – gives a path towards understanding interactions

Sentiment

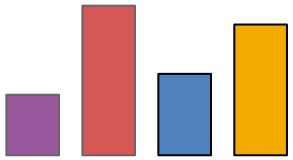


$R\ U_\ell U_{av}\ S$

Language/Agreement

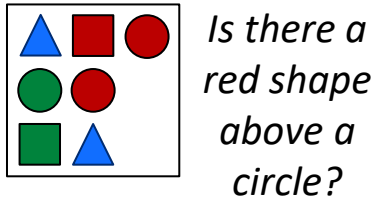


Sarcasm

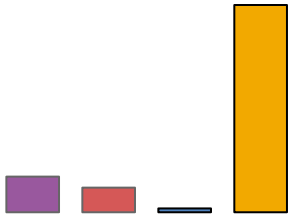


$R\ U_\ell U_{av}\ S$

Multimodal Transformer



VQA



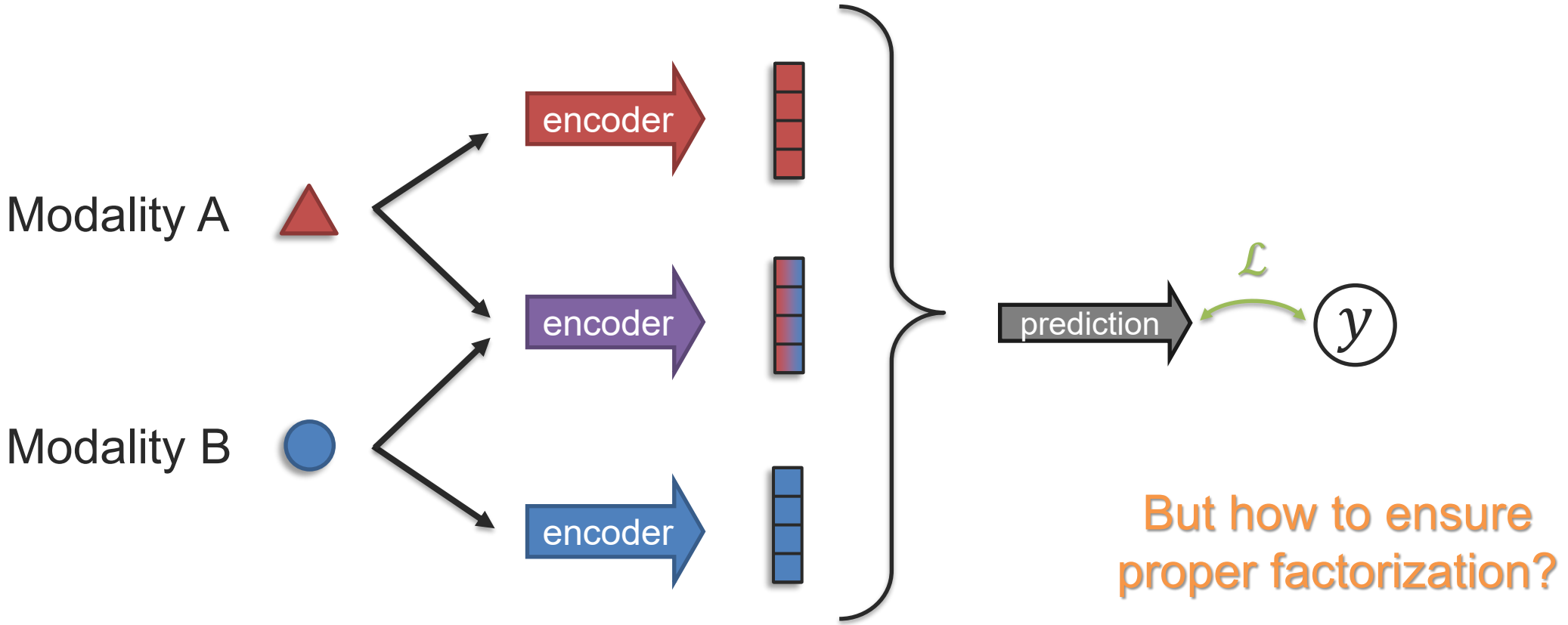
$R\ U_\ell\ U_i\ S$

Multiplicative/Transformer

Also matches human judgment of interactions, and other sanity checks on synthetic datasets
They can also be used to choose most appropriate models.

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

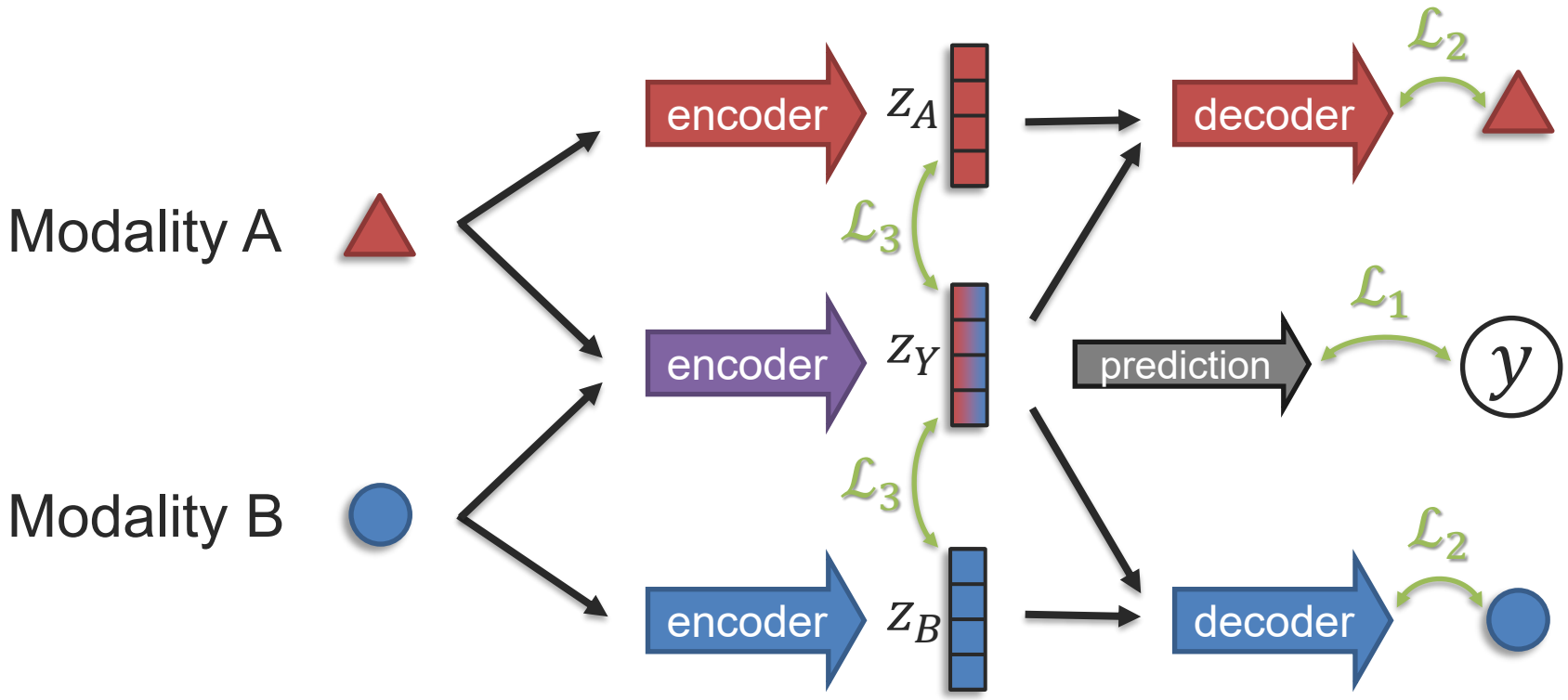
Factorized Multimodal Representations



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

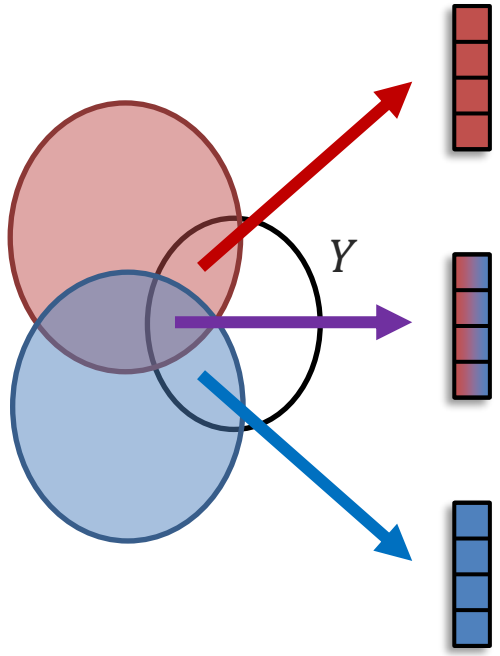
$\mathcal{L}_3$: no overlap

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



2 Maximize task-relevant **unique** information

$I(Z; Y|\bigcirc)$

1 Maximize task-relevant **shared** information

$I(Z; \bigcirc; Y)$ and $I(Z; \triangle; Y)$

3 Maximize task-relevant **unique** information

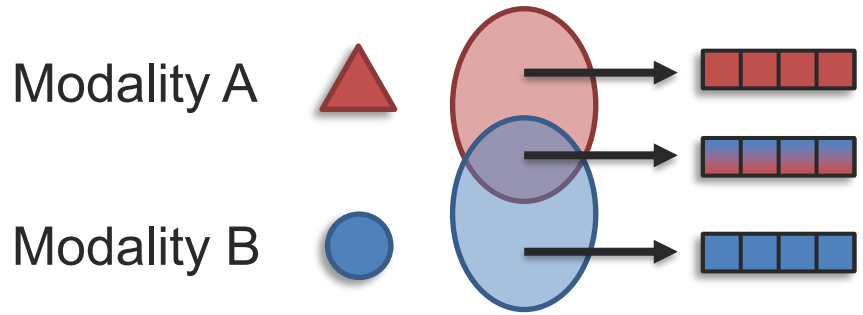
$I(Z; Y|\triangle)$

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

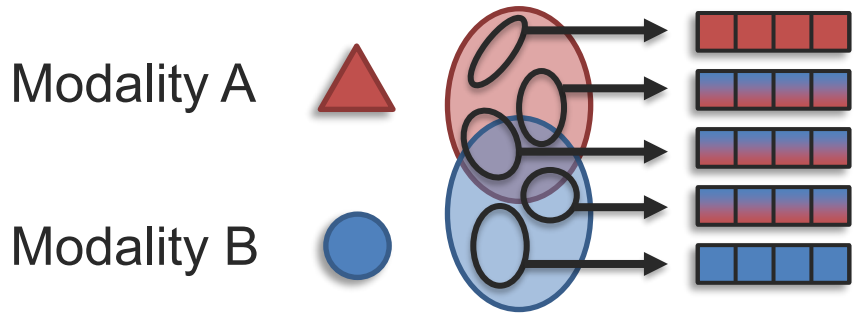
Fine-Grained Fission

How to automatically discover these internal clusters, factors?

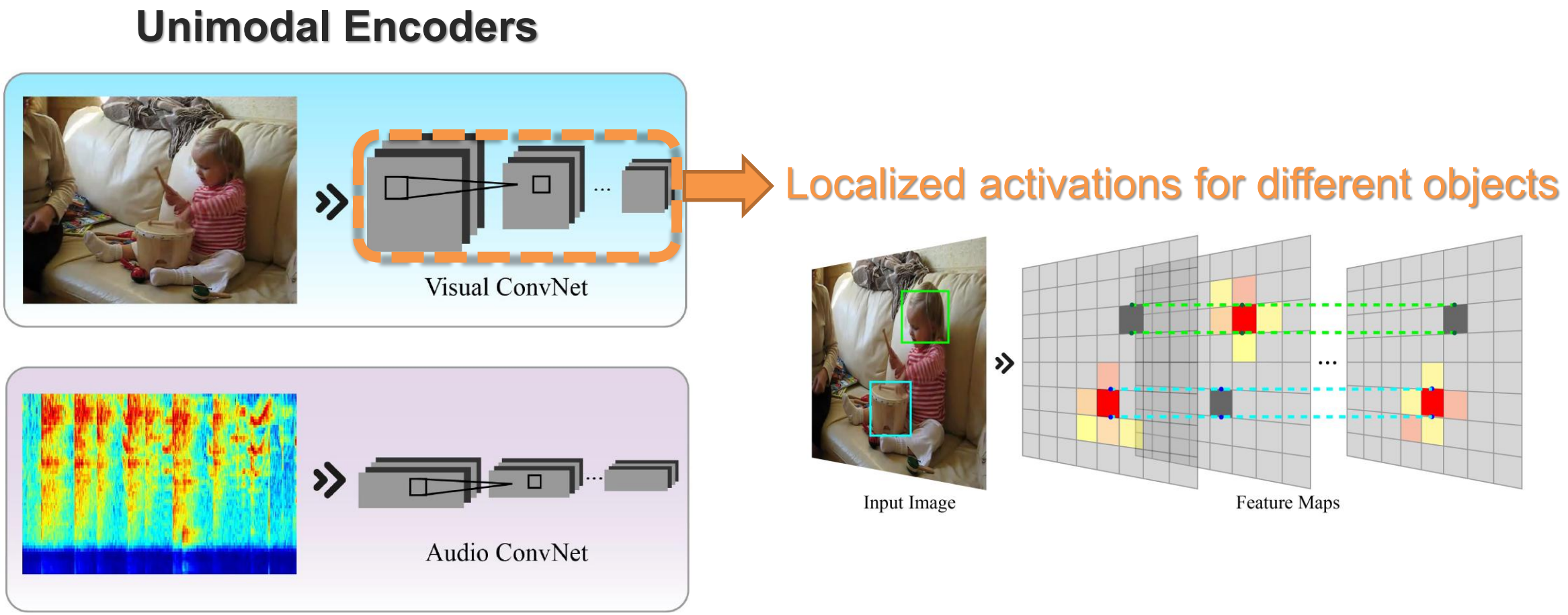
Modality-level fission:



Fine-grained fission:



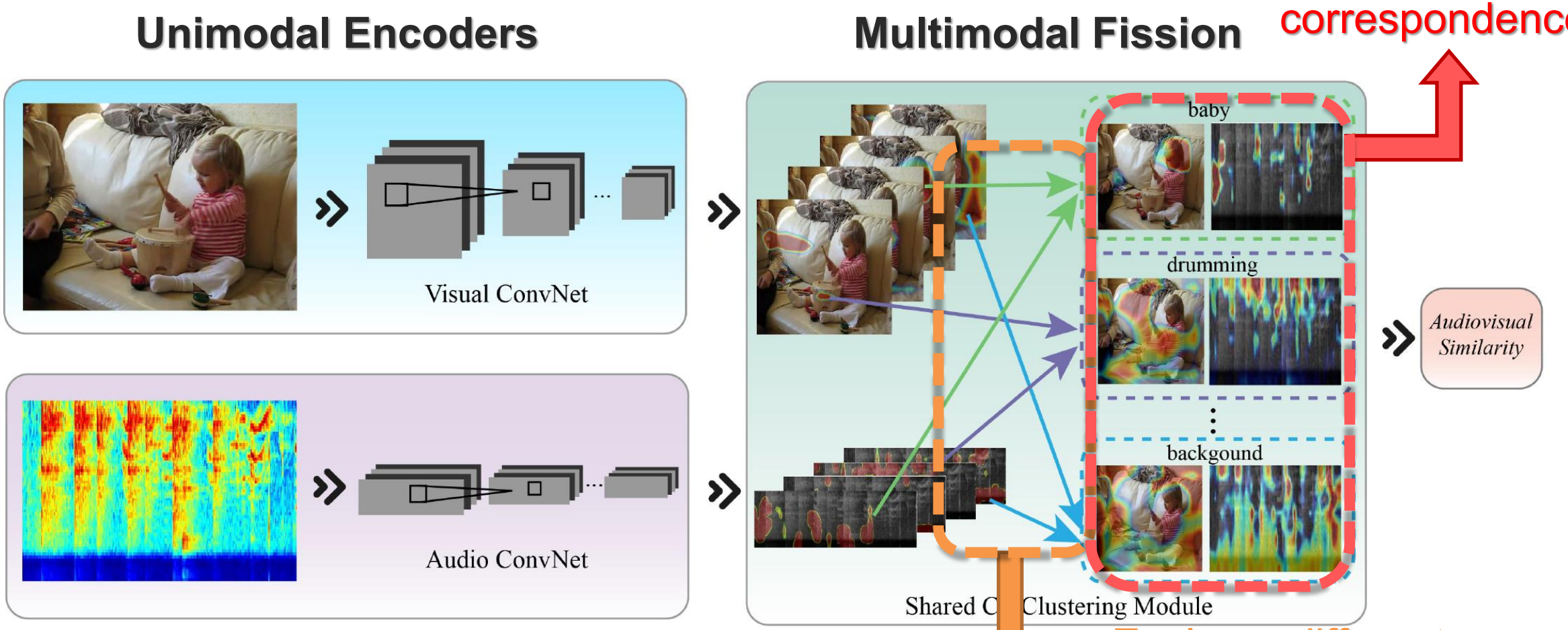
Fine-Grained Fission – A Clustering Approach



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