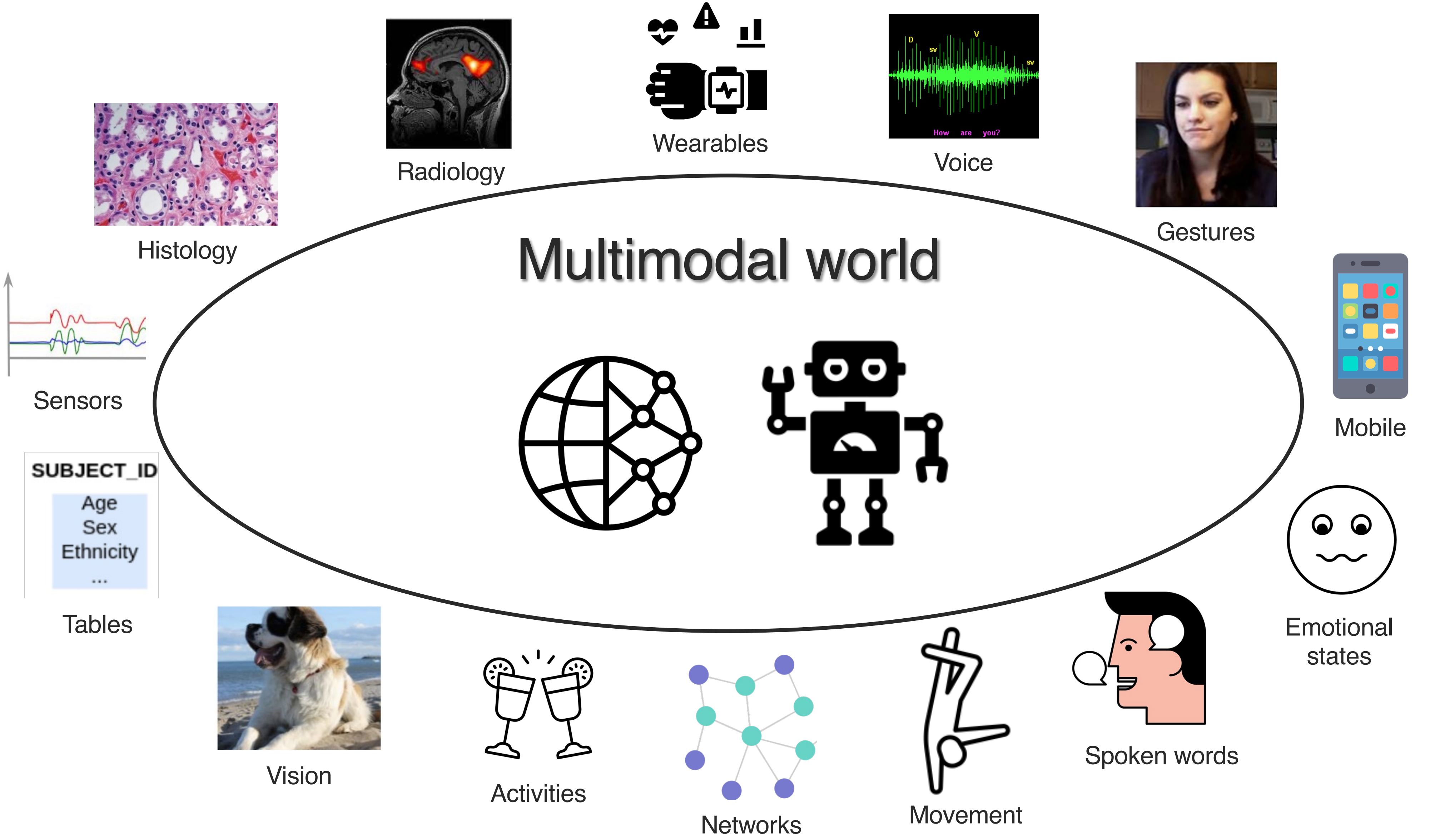
Multimodal Deep Learning

Paul Liang
Machine Learning Department
Carnegie Mellon University

https://www.cs.cmu.edu/~pliang/
pliang@cs.cmu.edu
https://github.com/pliang279
@pliang279

Multimodal Artificial Intelligence



Multimodal Human Communication

Understanding human language and gestures

Diverse annotations

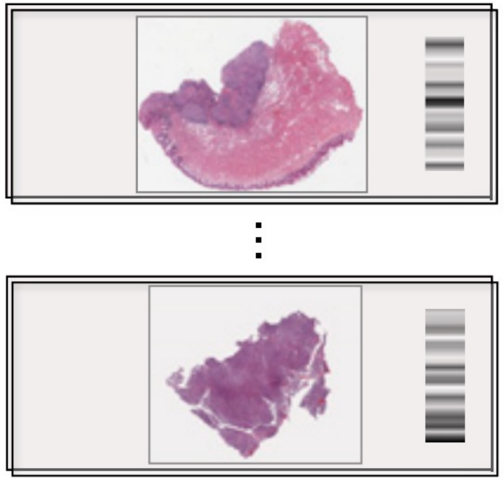
- Sentiment
- Emotions
- Sarcasm
- Humor
- Personalities

Practical applications

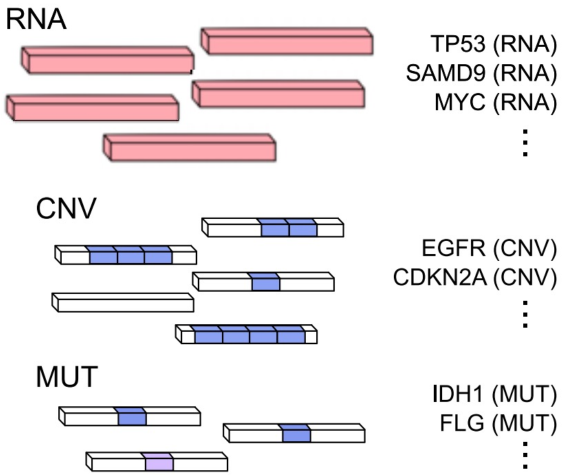
Healthcare Modalities

Medicine and healthcare

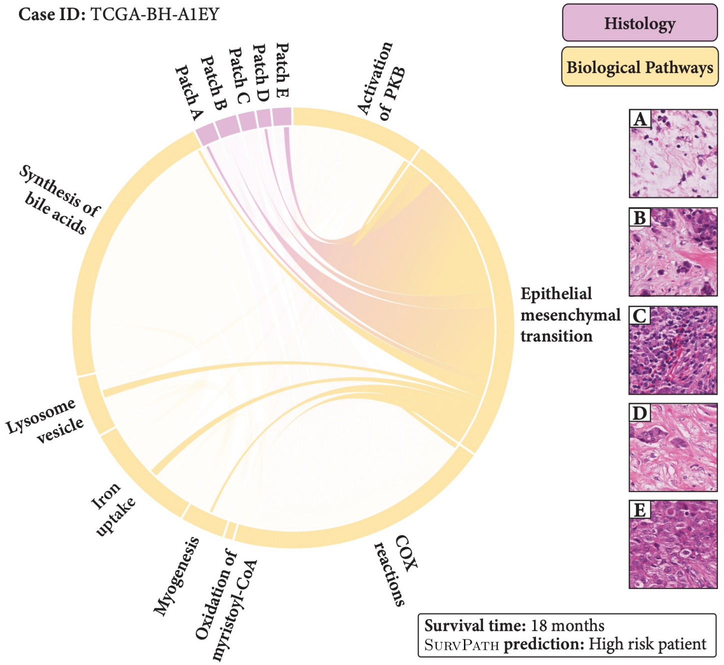
Histology images



Genomics profile



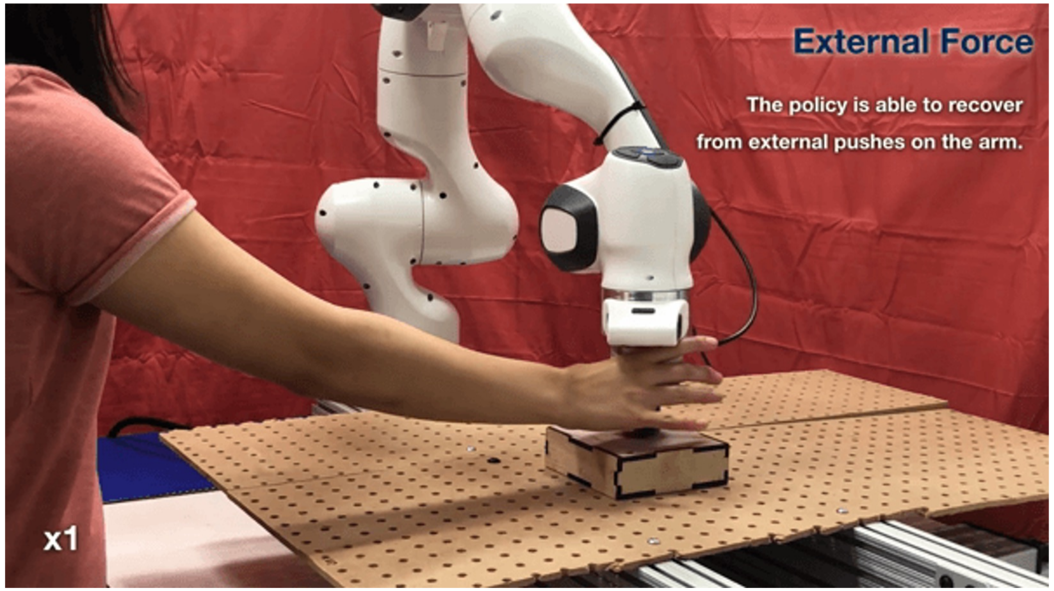
Transcriptomics

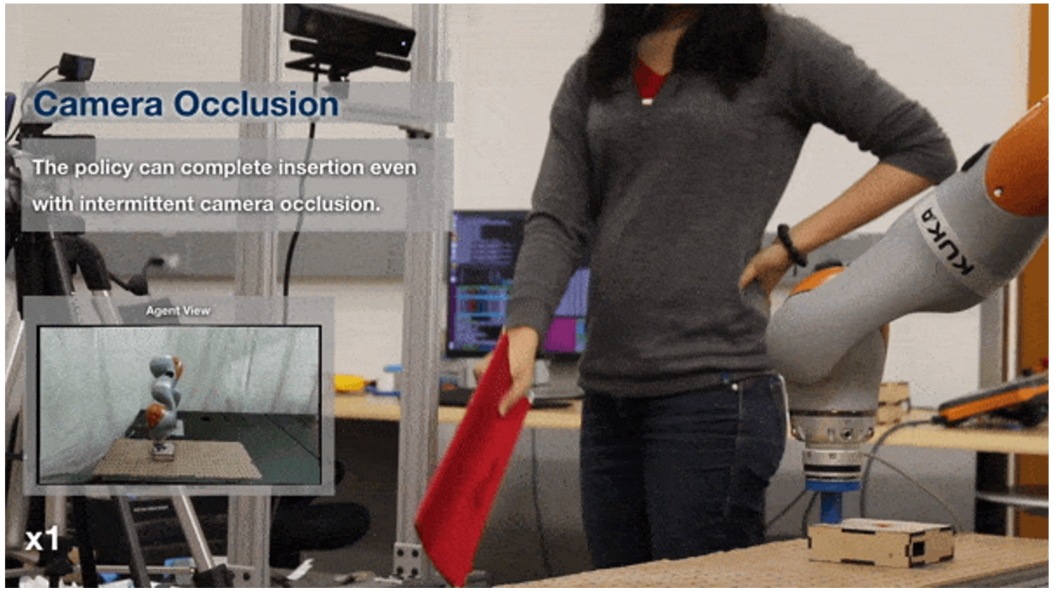


[Jaume et al., Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction. 2023]
[Liang et al., Quantifying & Modeling Multimodal Interactions: An Information Decomposition Framework. NeurIPS 2023]

Multisensory Robotic Intelligence

Multisensor fusion in robotics





+ robustness

[Lee et al., Making Sense of Vision and Touch: Self-Supervised Learning of Multimodal Representations for Contact-Rich Tasks. ICRA 2019]
[Liang et al., MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. NeurIPS 2021]

Multimodal Machine Learning – Surveys, Tutorials and Courses

Foundations and Recent Trends in Multimodal Machine Learning

Paul Liang, Amir Zadeh and Louis-Philippe Morency

- ☑ 6 core challenges
- ☑ 50+ taxonomic classes
- ☑ 700+ referenced papers

https://arxiv.org/abs/2209.03430

Tutorials: ICML 2023, CVPR 2022, NAACL 2022

Graduate-level courses:

Multimodal Machine Learning (12th edition)

https://cmu-multicomp-lab.github.io/mmml-course/fall2022/

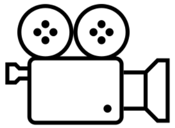
Advanced Topics in Multimodal ML

https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2023/

What is a Modality?

Definition

Modality refers to the way in which something expressed or perceived.



from a sensor

Raw Modalities
(closest from sensor)

Abstract Modalities
(farthest from sensor)

Examples:

Speech signal

Language

Sentiment intensity

Image

Detected objects

Object categories

[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

Multimodal **is the scientific study of**

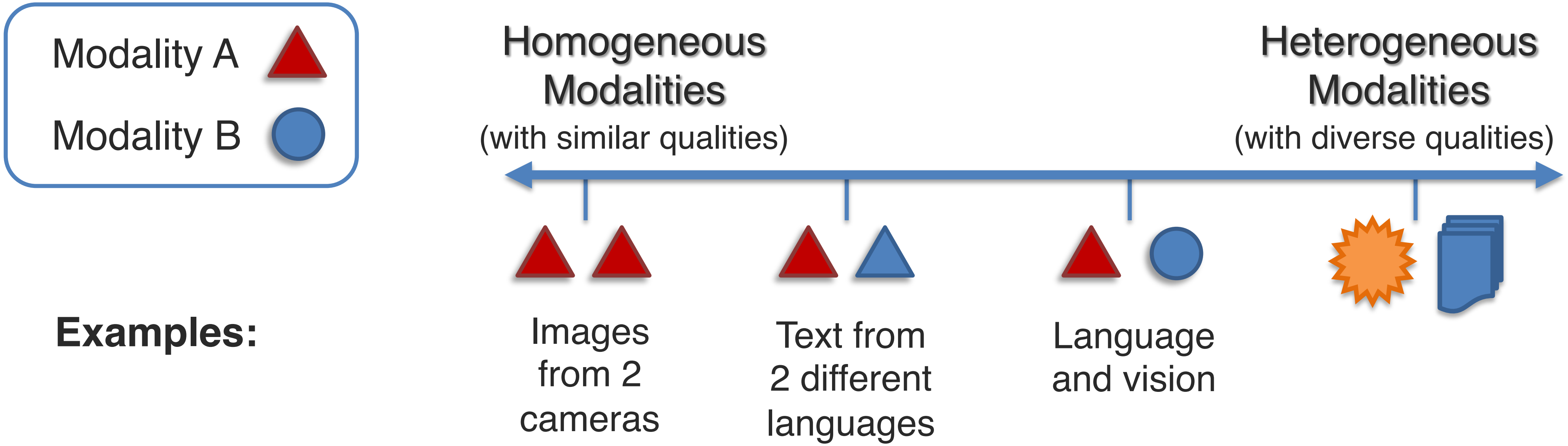
heterogeneous and interconnected data

$\underbrace{\text{interconnected}}_{\text{Connected + Interacting}}$

[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Heterogeneous Modalities

Heterogeneous: Diverse qualities, structures and representations.



Abstract modalities are more likely to be homogeneous

[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Dimensions of Heterogeneity

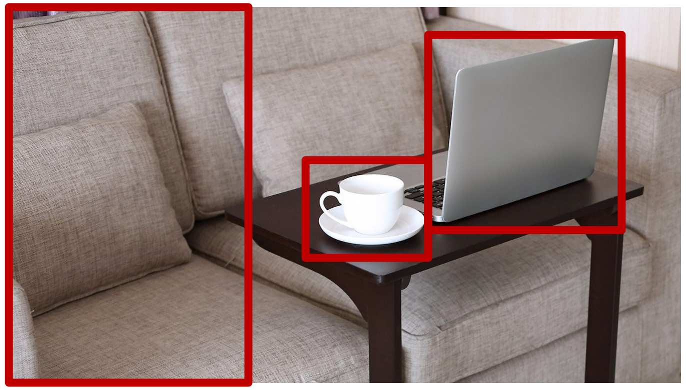
Information present in different modalities will often show diverse qualities, structures, and representations.

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

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



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

1 **Distribution:** discrete or continuous, support

{teacup, right, laptop, clean, room}

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

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

2 **Granularity:** sampling rate and frequency

objects per image

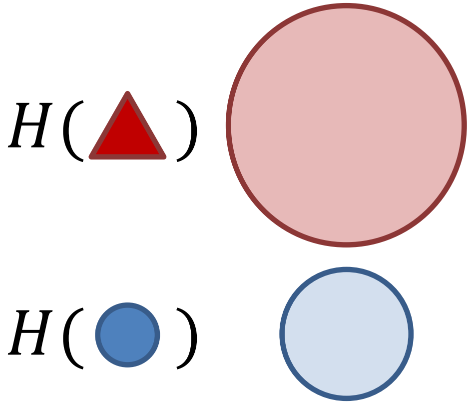
words per minute

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

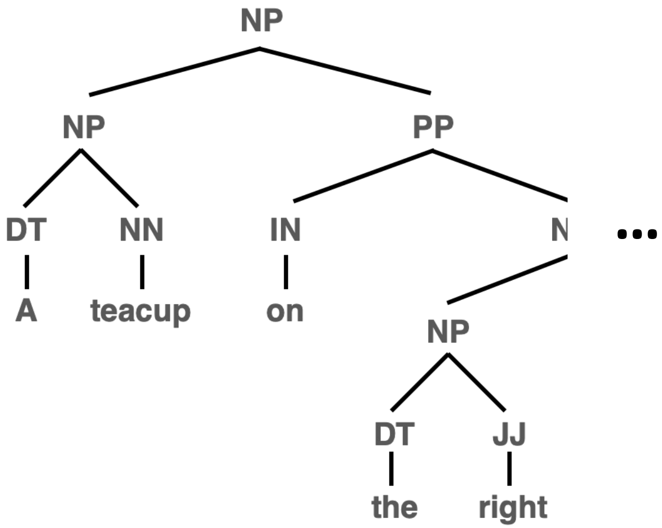
A teacup on the right of a laptop in a clean room.

3 **Information:** entropy and density

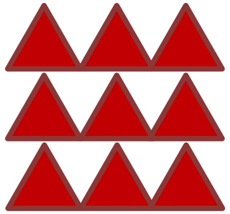


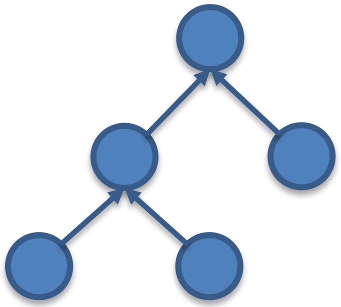
Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



4 **Structure:** static, temporal, spatial, hierarchical





Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

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

5 **Noise:** uncertainty, signal-to-noise ratio, missing data

teacup → teacip

right → rihjt

Connected Modalities

Connected: Shared information that relates modalities

[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

Connected Modalities

Connected: Shared information that relates modalities

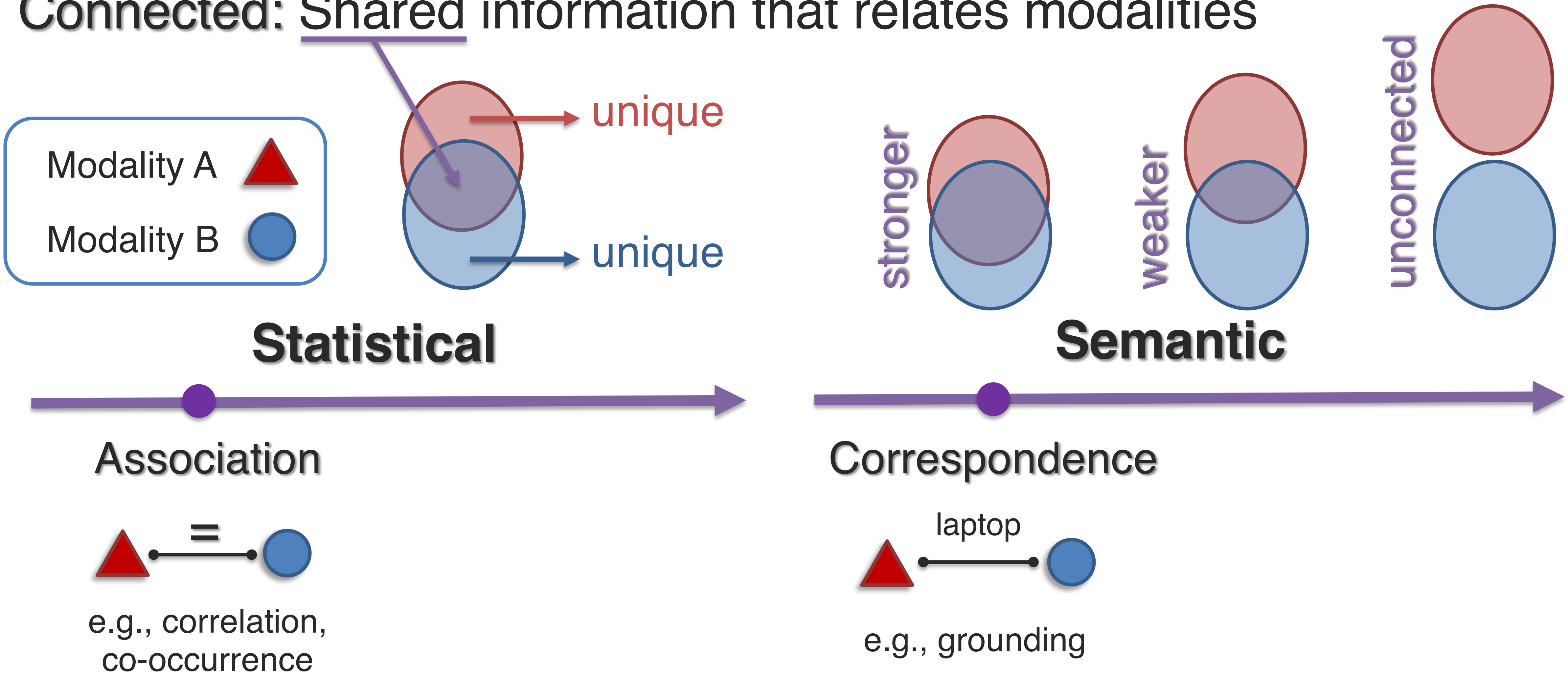
A teacup on the right of a laptop in a clean room.

↔ ***teacup***

↔ ***laptop***

Connected Modalities

Connected: Shared information that relates modalities



[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

Connected Modalities

Connected: Shared information that relates modalities

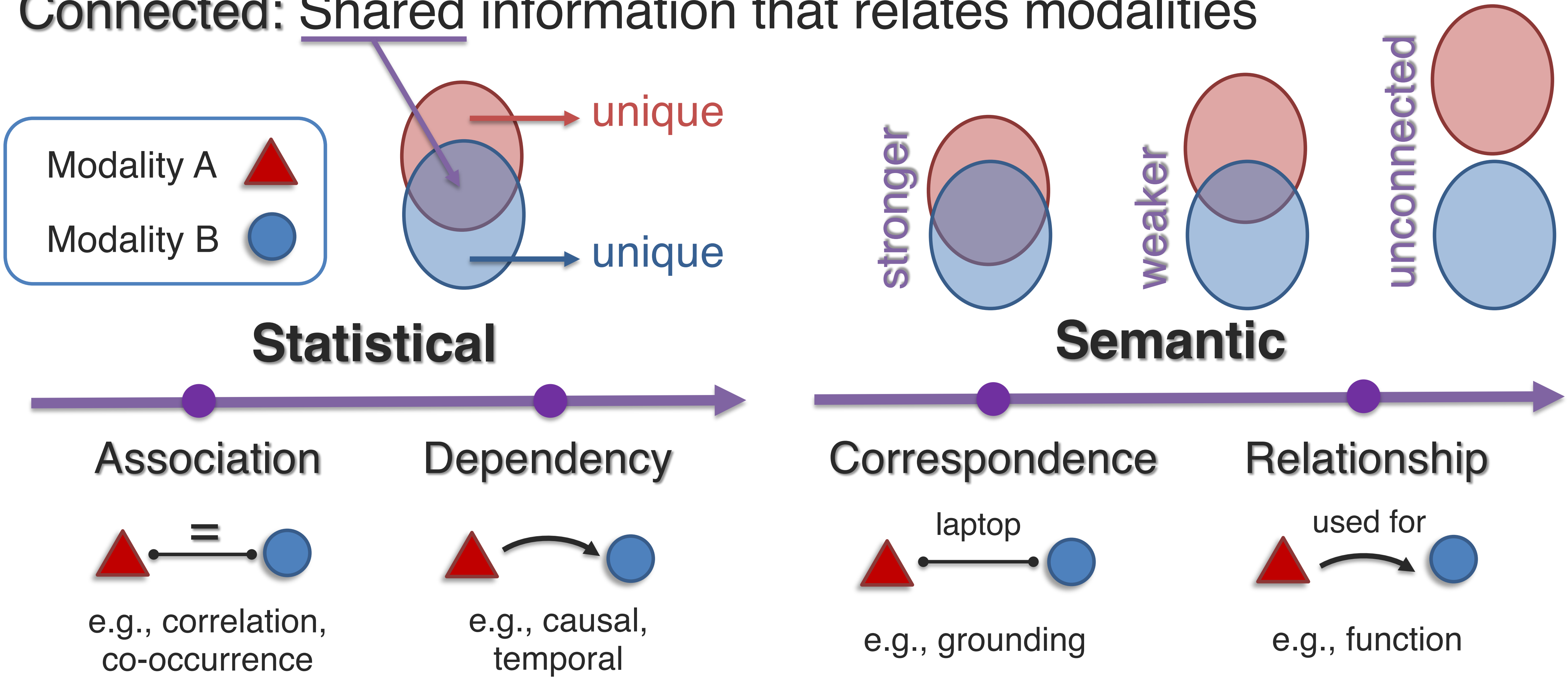
A teacup on the right of a laptop in a clean room.

↔ **clean**

↔ **room**

Connected Modalities

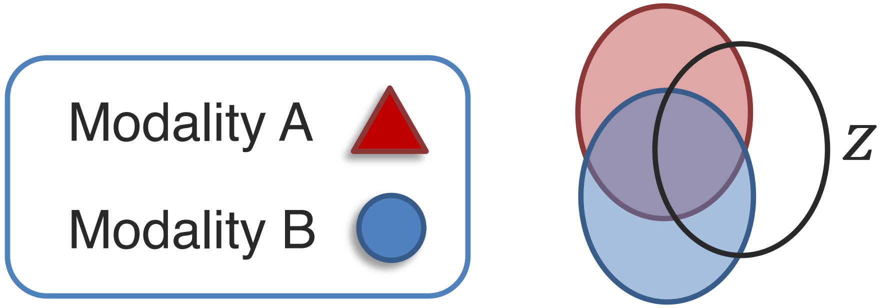
Connected: Shared information that relates modalities

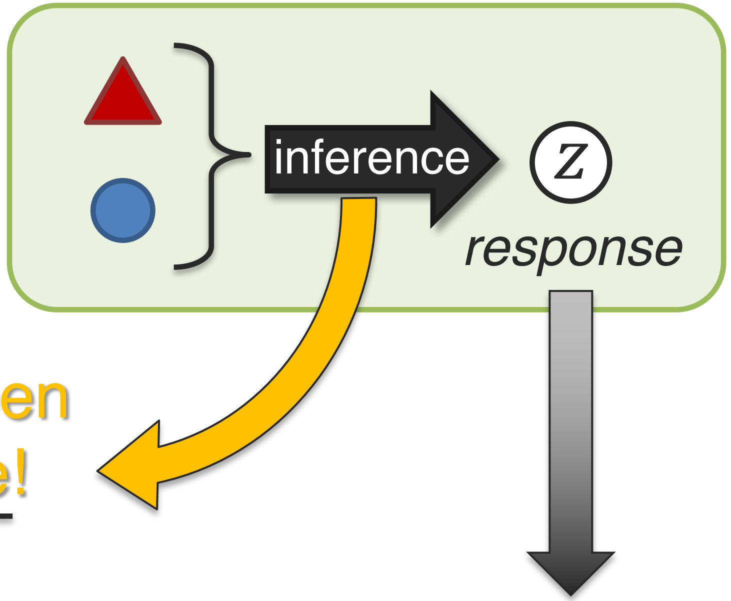


[Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. ICML 2023, CVPR 2022, NAACL 2022 Tutorials]

Interacting Modalities

Interacting: process affecting each modality, creating new response





Interactions happen during inference!

"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation

representation

$\hat{y}$ *prediction*

modality C

Interacting Modalities

Interactions: How multimodal information changes when modalities are combined for a response.

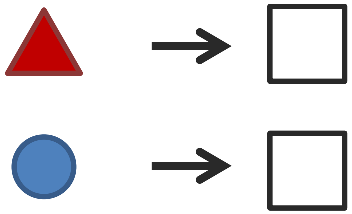
Is this indoors?

inference → *Yes!*

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

inference → *Yes!*

Redundant



Interacting Modalities

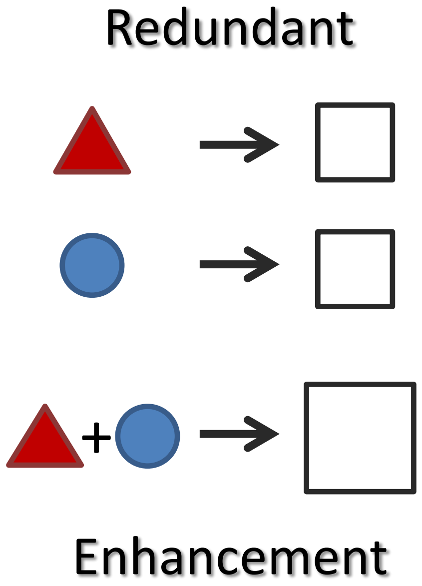
Interactions: How multimodal information changes when modalities are combined for a response.

Is this indoors?

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

inference

Yes!



Interacting Modalities

Interactions: How multimodal information changes when modalities are combined for a response.

Is this a living room?

inference

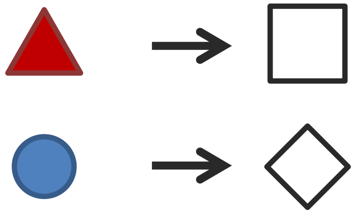
Yes!

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

inference

No, probably study room.

Non-redundant



Interacting Modalities

Interactions: How multimodal information changes when modalities are combined for a response.

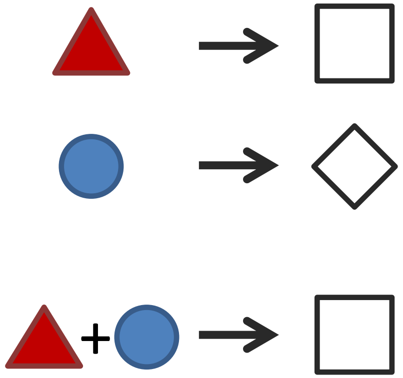
Is this a living room?

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

inference

Yes!

Non-redundant



Dominance

Interacting Modalities

Interactions: How multimodal information changes when modalities are combined for a response.

Should I work here?

inference

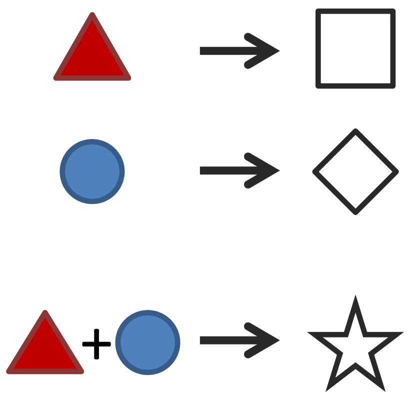
Maybe? Comfy sofa but table's too small.

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

inference

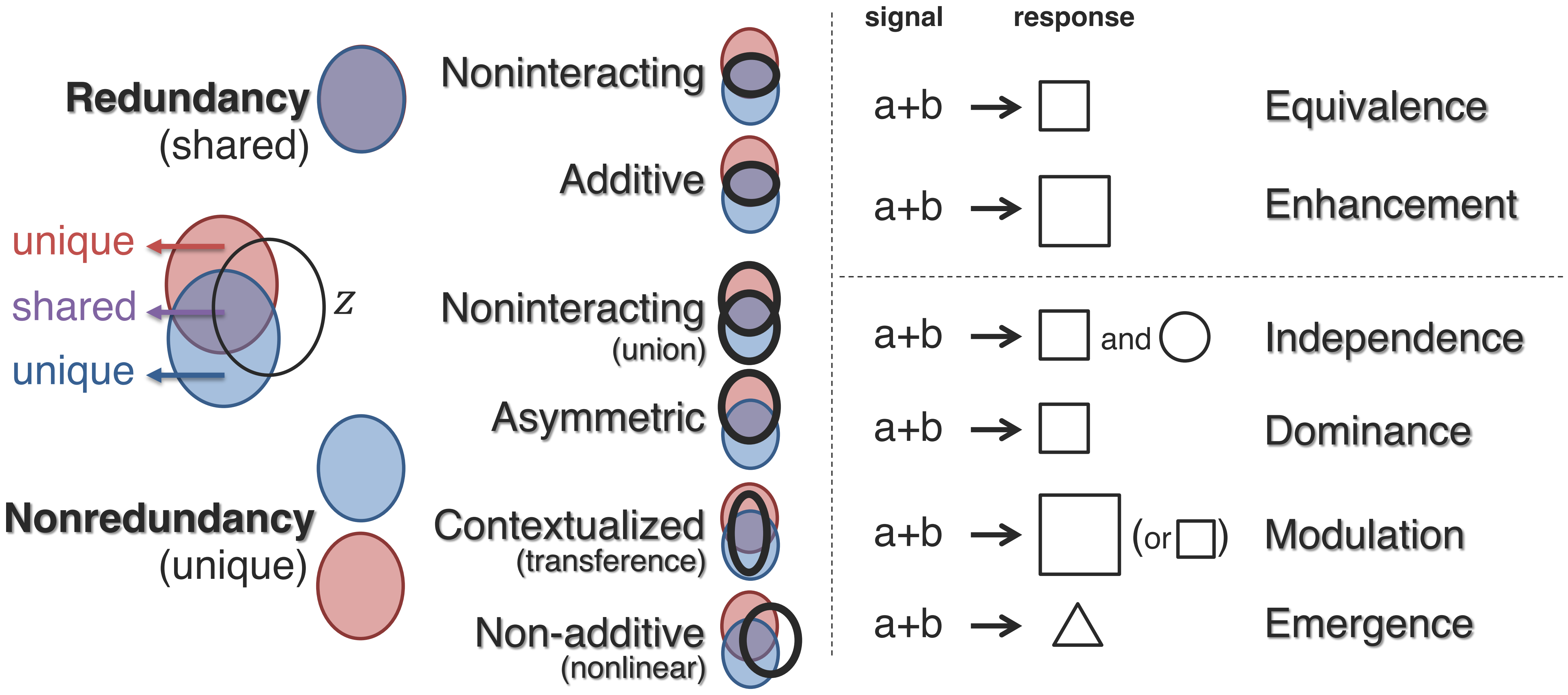
Maybe? Clean and there's tea.

Non-redundant



Emergence

Cross-modal Interaction Mechanics



What is Multimodal? → *Why is it hard?* → *What is next?*

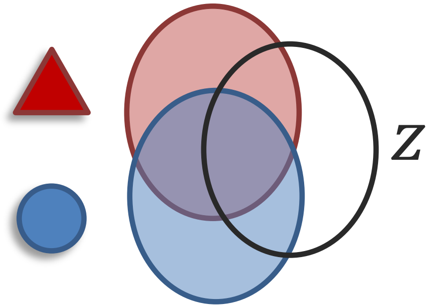
Heterogeneous

+

Connected

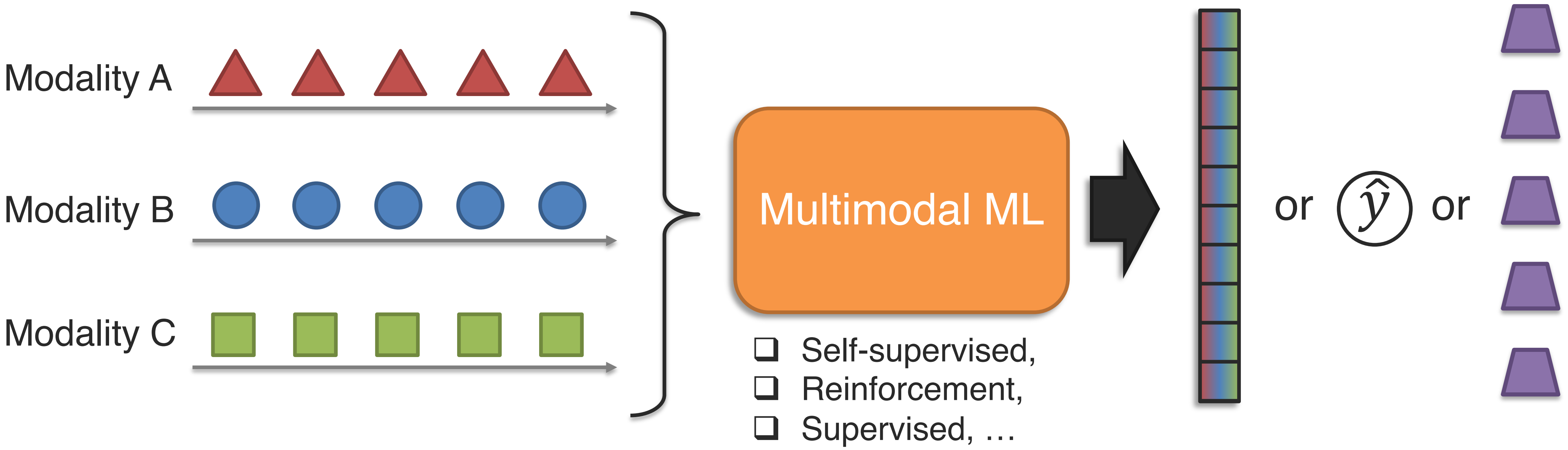
+

Interacting



Multimodal is the scientific study of **heterogeneous** and **interconnected** data ☺

Multimodal Machine Learning



What are the core multimodal technical challenges, understudied in conventional machine learning?

Challenge 1: Representation

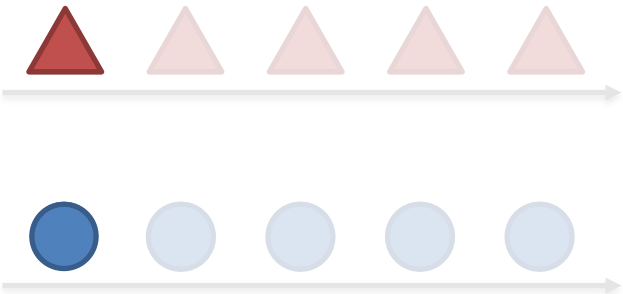
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

➡ 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

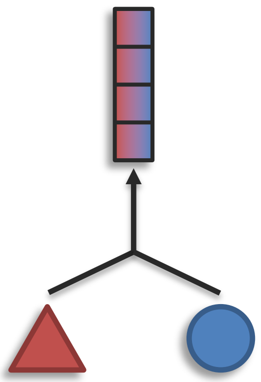
[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Challenge 1: Representation

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

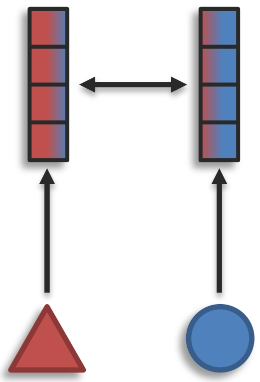
Sub-challenges:

Fusion



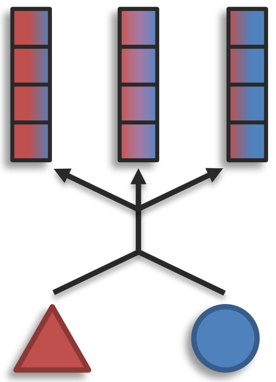
modalities **>** # representations

Coordination



modalities **=** # representations

Fission



modalities **<** # representations

[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Challenge 2: Alignment

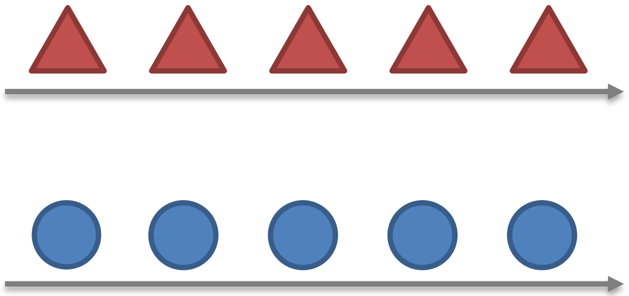
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

➡ Most modalities have internal structure with multiple elements

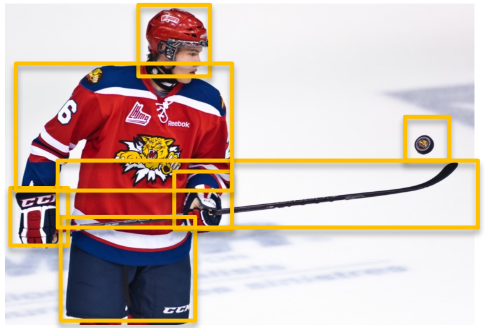
Elements with temporal structure:

Modality A

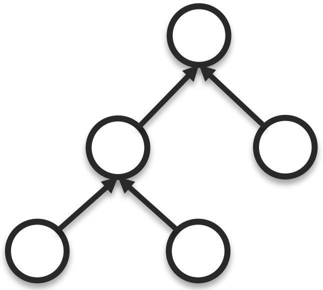
Modality B



Other structured examples:



Spatial



Hierarchical

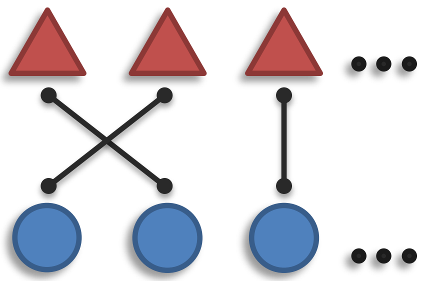
[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Challenge 2: Alignment

Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

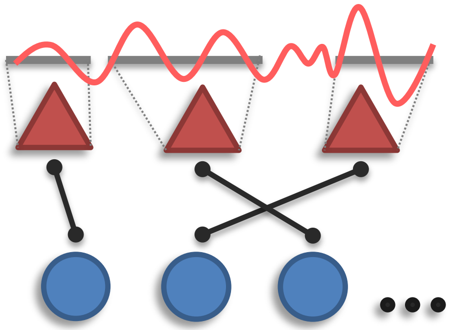
Sub-challenges:

Discrete Alignment



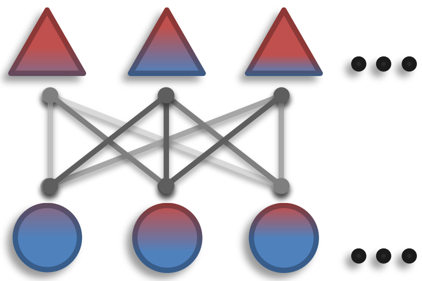
Discrete elements and connections

Continuous Alignment



Segmentation and continuous warping

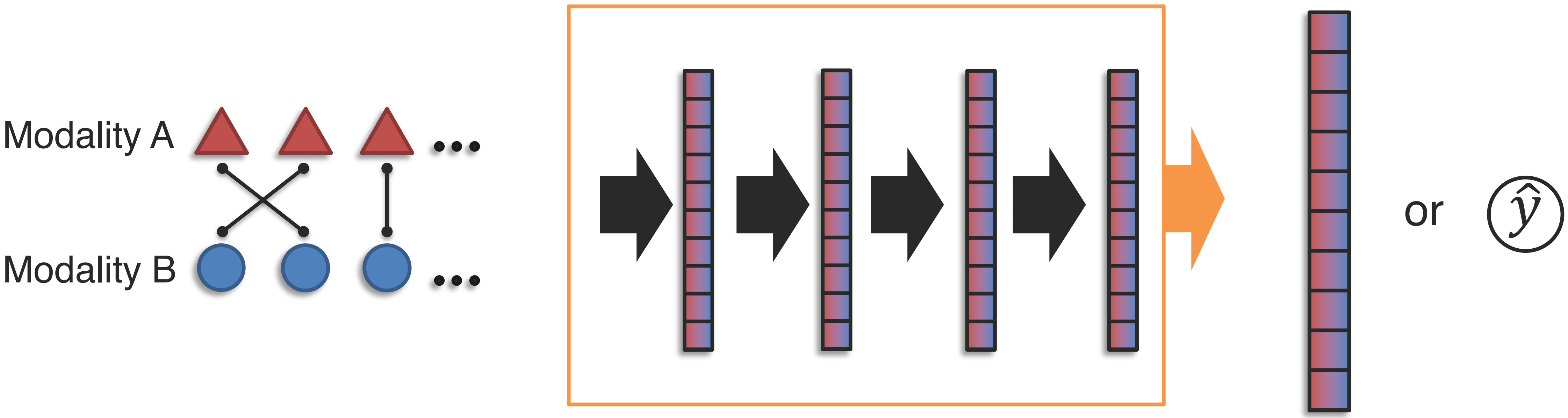
Contextualized Representation



Alignment + representation

Challenge 3: Reasoning

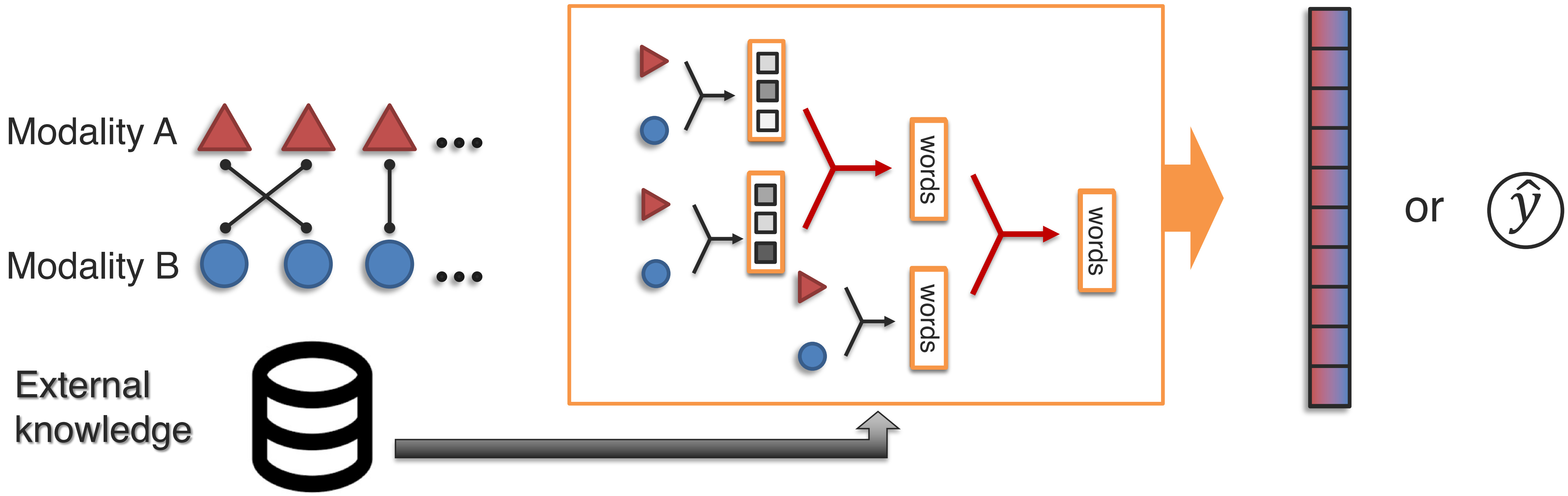
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Challenge 3: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

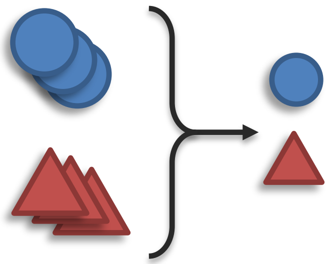


Challenge 4: Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence

Sub-challenges:

Summarization



Reduction

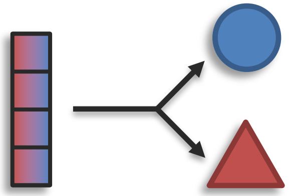
□ > □

Translation

Maintenance

□ = □

Creation



Expansion

□ < □

Information:
(content)

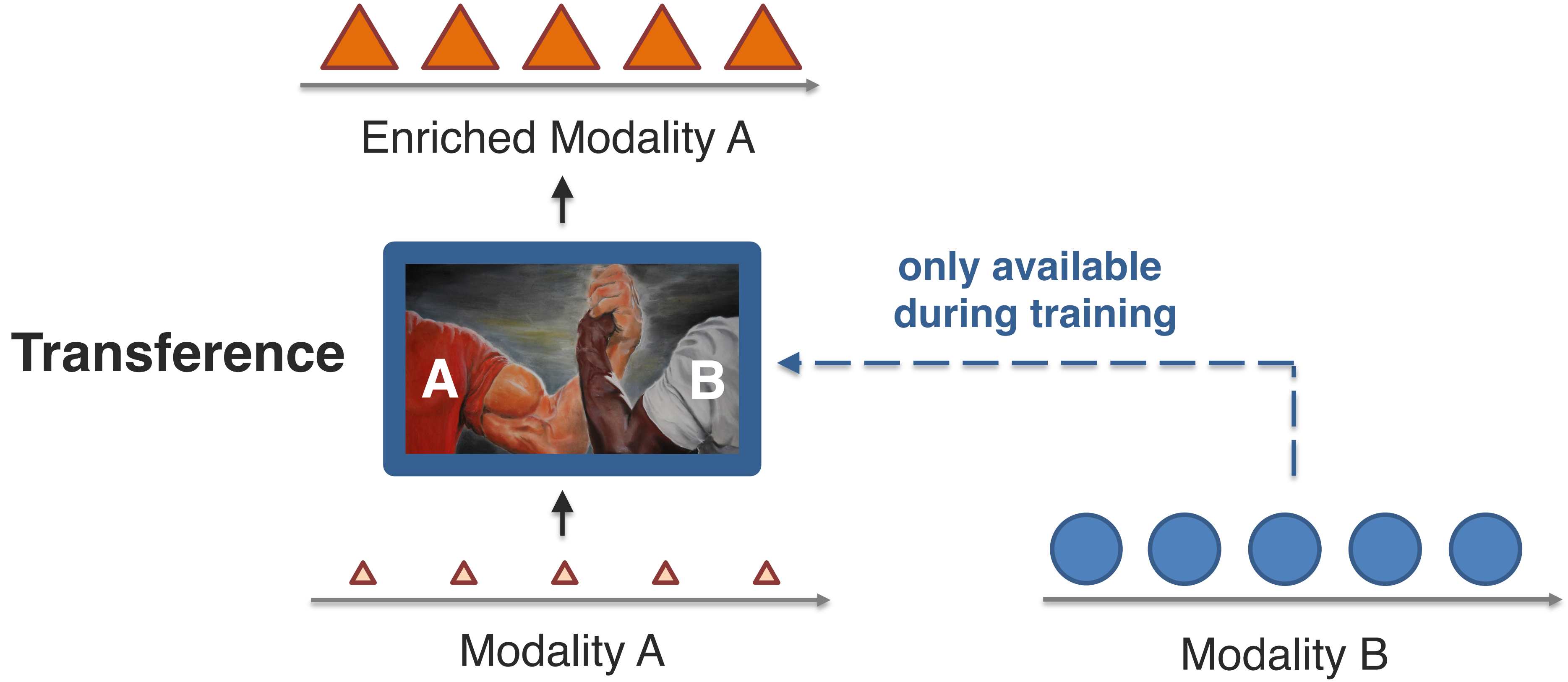
Challenge 4: Generation

An astronaut riding a horse in the style of Andy Warhol.

A bowl of soup that is a portal to another dimension as digital art

Challenge 5: Transference

Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources

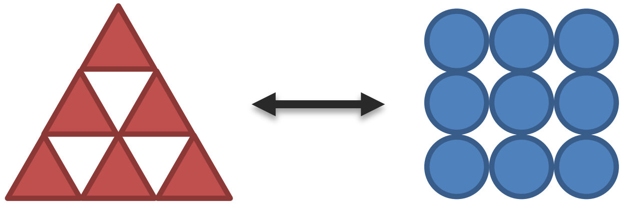


Challenge 6: Quantification

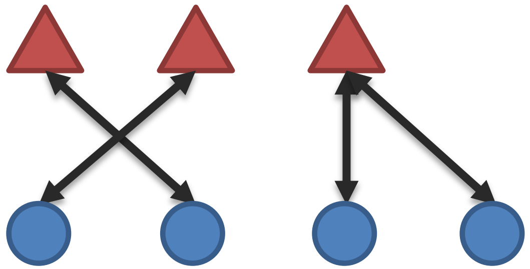
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions and the multimodal learning process

Sub-challenges:

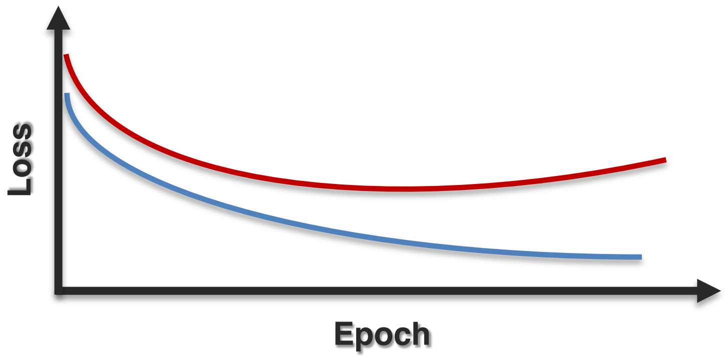
Heterogeneity



Connections & Interactions

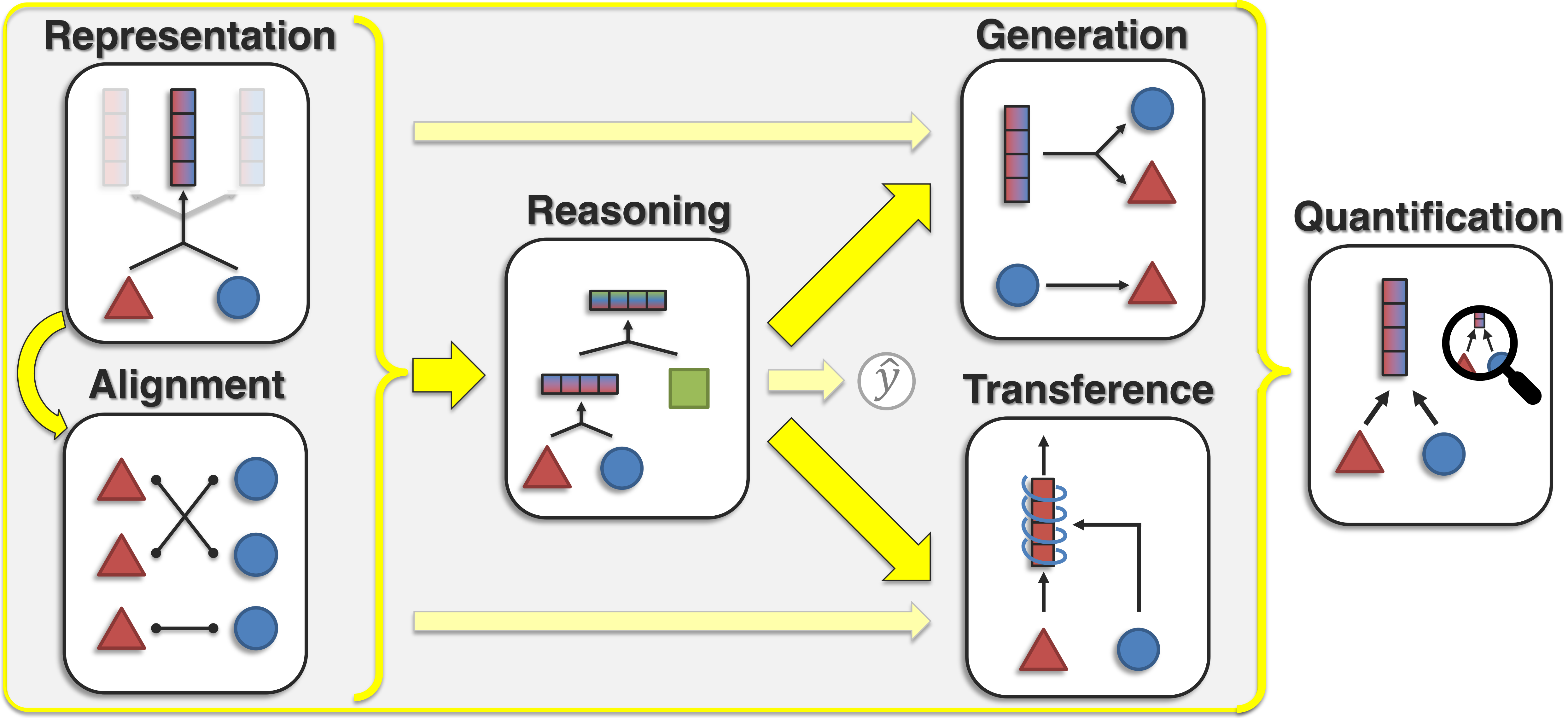


Learning

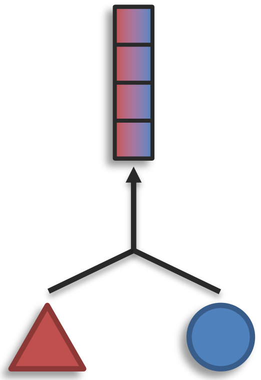


[Liang, Zadeh, Morency, Foundations and Trends in Multimodal Machine Learning. Tutorials at ICML 2023, CVPR 2022, NAACL 2022]

Core Multimodal Challenges

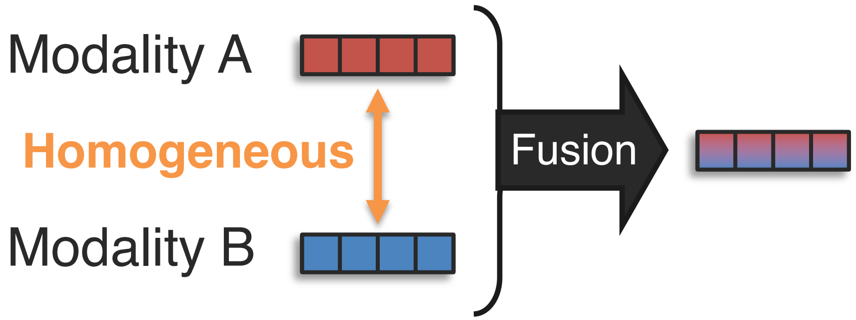


Sub-Challenge: Representation Fusion

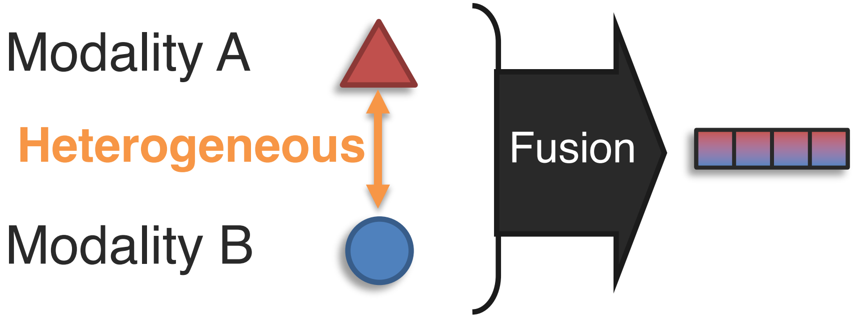


Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Basic fusion:

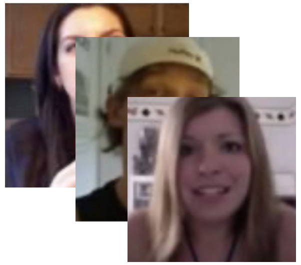


Raw-modality fusion:



From Additive to Multiplicative

300 book reviews



y: audience score

x_A: percentage of smiling

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

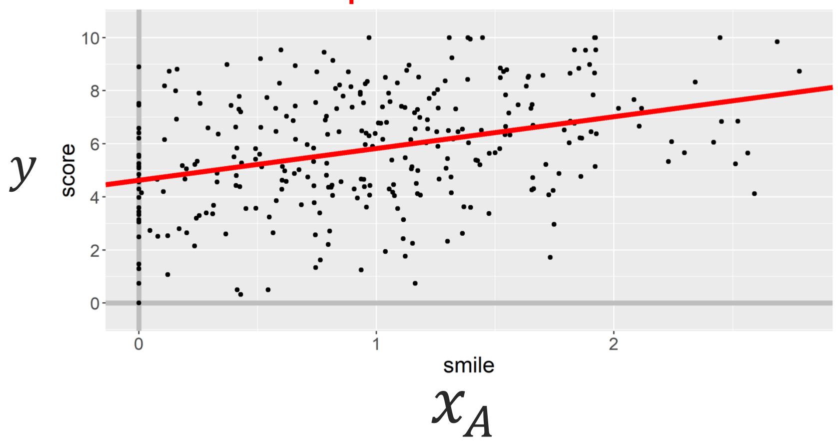
H1: *Does smiling reveal what the audience score was?*

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

Linear regression:

$$y = w_0 + \boxed{w_1} x_A + \epsilon$$

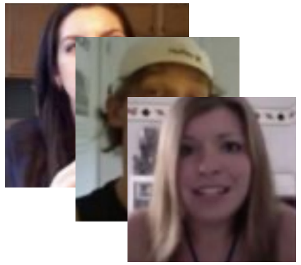
slope



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

From Additive to Multiplicative

300 book reviews



y: audience score

x_A: percentage of smiling

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

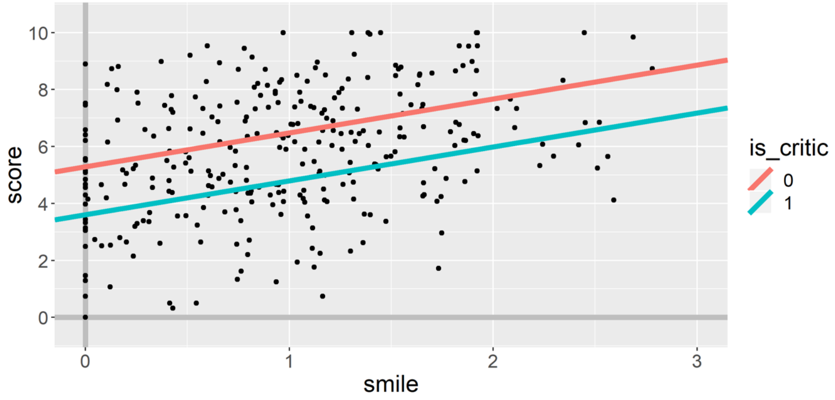
H1: *Does smiling reveal what the audience score was?*

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

Linear regression:

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

y



x_A

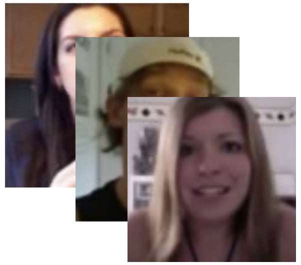
	Estimate	95% CI
w_0	5.29	$[4.86, 5.73]$
w_1	1.19	$[0.85, 1.53]$
w_2	-1.69	$[-2.14, -1.24]$

Positive effect

Negative effect

From Additive to Multiplicative

300 book reviews



y: audience score

x_A: percentage of smiling

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

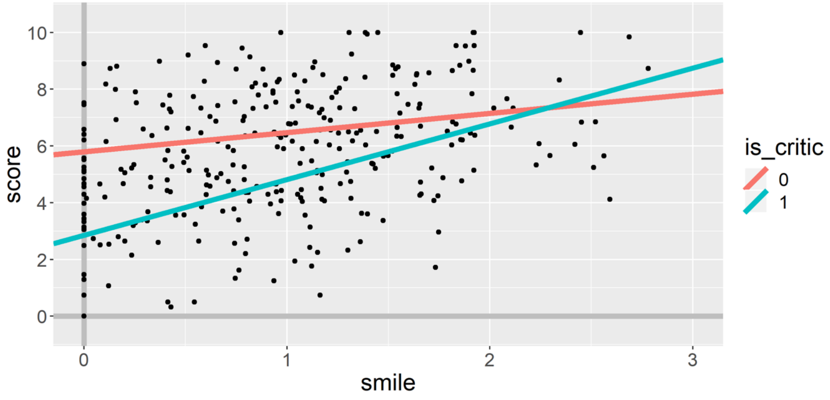
***H1:** Does smiling reveal what the audience score was?*

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

Linear regression:

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

y

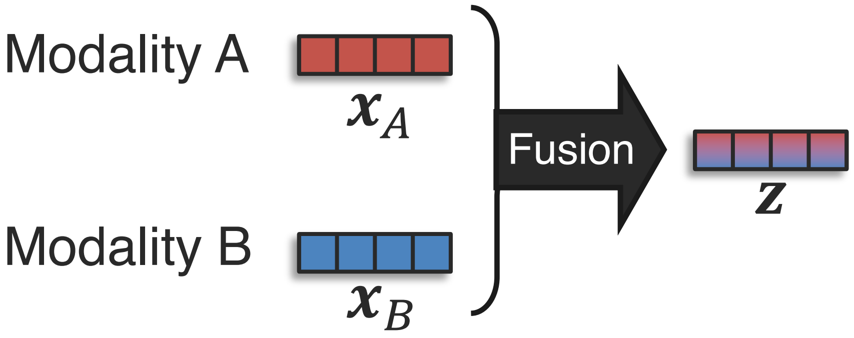


x_A

	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 Fusion – Additive Interactions

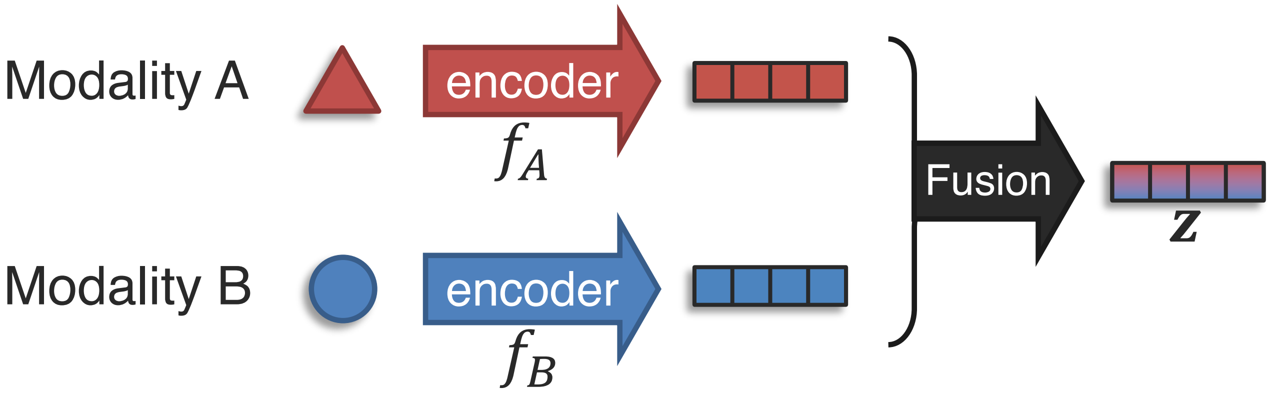


Additive fusion:

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

1-layer neural network can be seen as additive

With unimodal encoders:

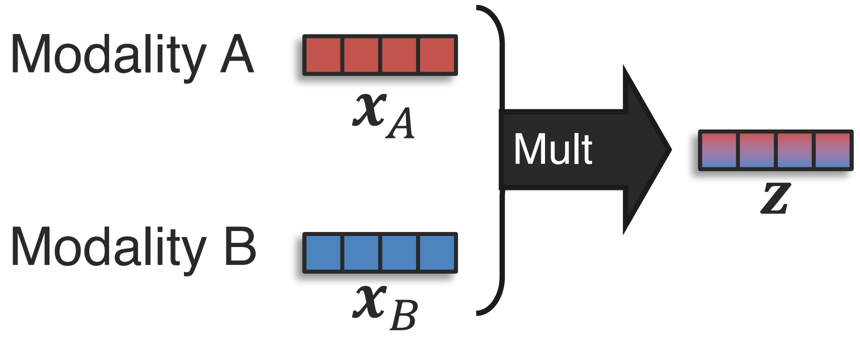


Additive fusion:

$$\boldsymbol{z} = f_A(\triangle) + f_B(\bigcirc)$$

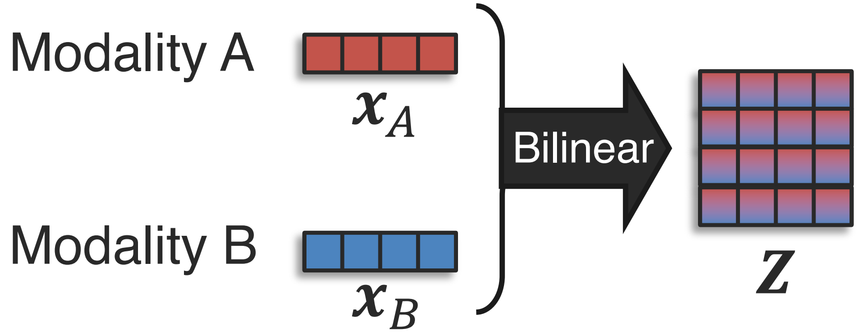
It could be seen as an ensemble approach
(late fusion)

Multiplicative Interactions



Simple multiplicative fusion:

$$\boldsymbol{z} = \boldsymbol{w}(\boldsymbol{x}_A \times \boldsymbol{x}_B)$$

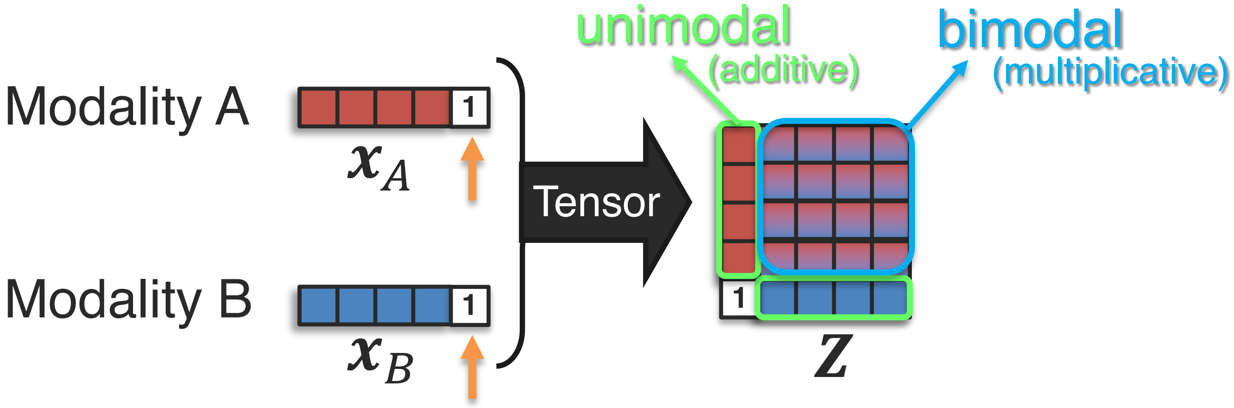


Bilinear Fusion:

$$\boldsymbol{Z} = \boldsymbol{W}(\boldsymbol{x}_A^T \cdot \boldsymbol{x}_B)$$

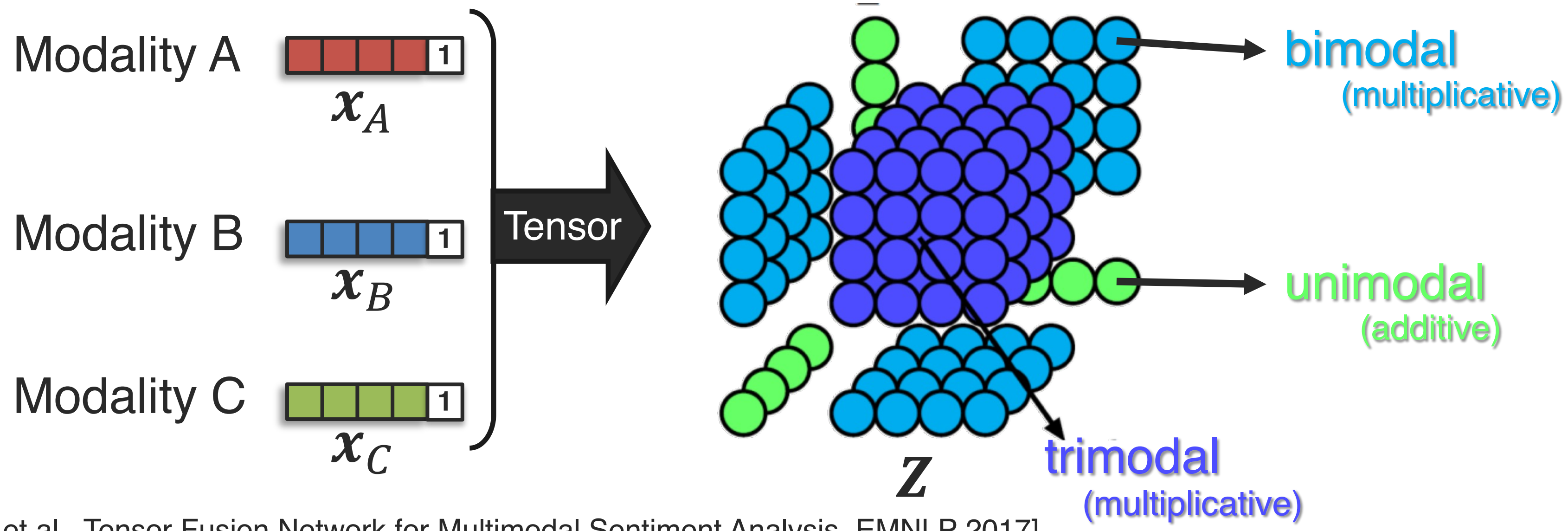
[Jayakumar et al., Multiplicative Interactions and Where to Find Them. ICLR 2020]

Tensor Fusion



Tensor Fusion (bimodal):

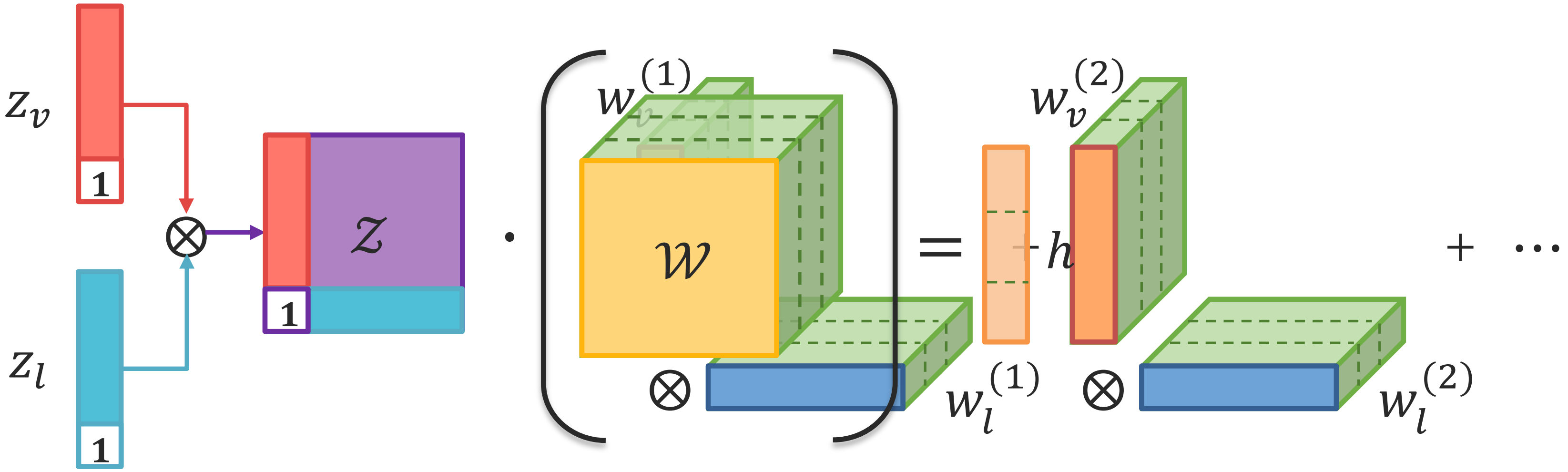
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

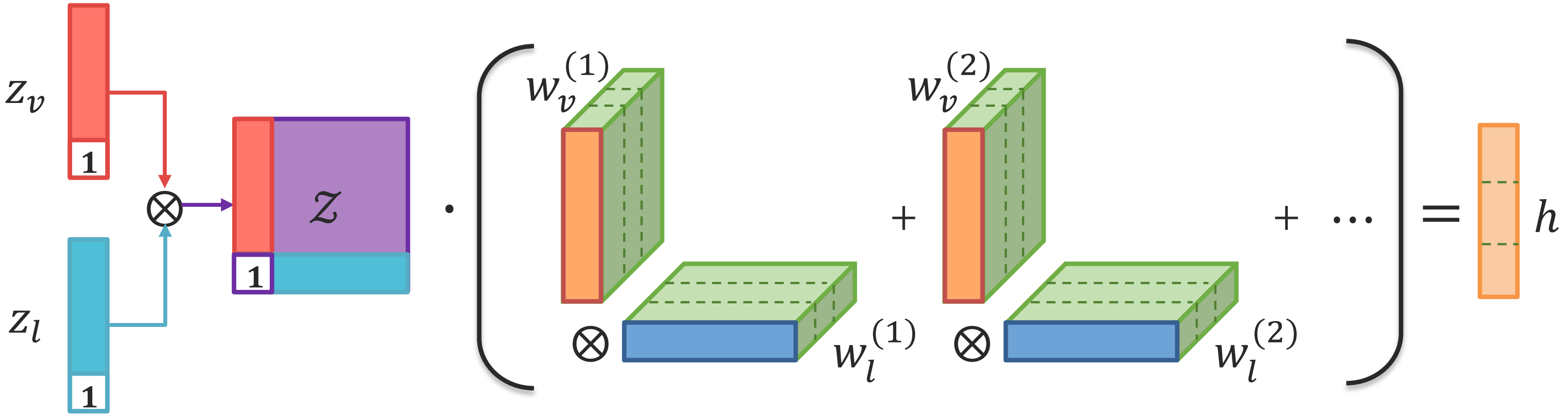
[Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017]
[Hou et al., Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling. NeurIPS 2019]

Low-rank Tensor Fusion



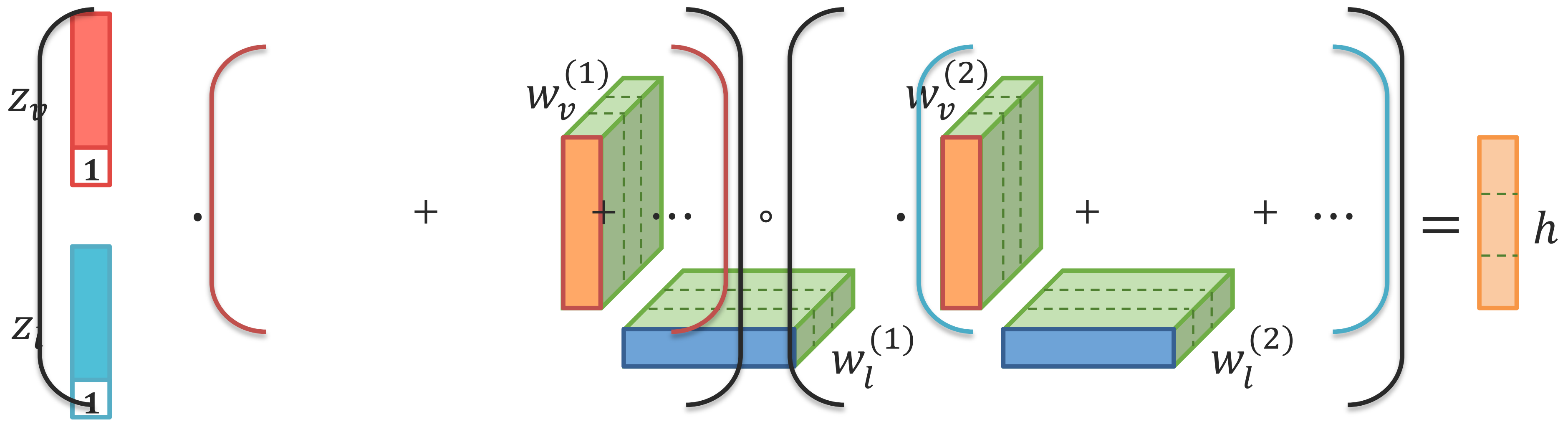
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Tensor Fusion



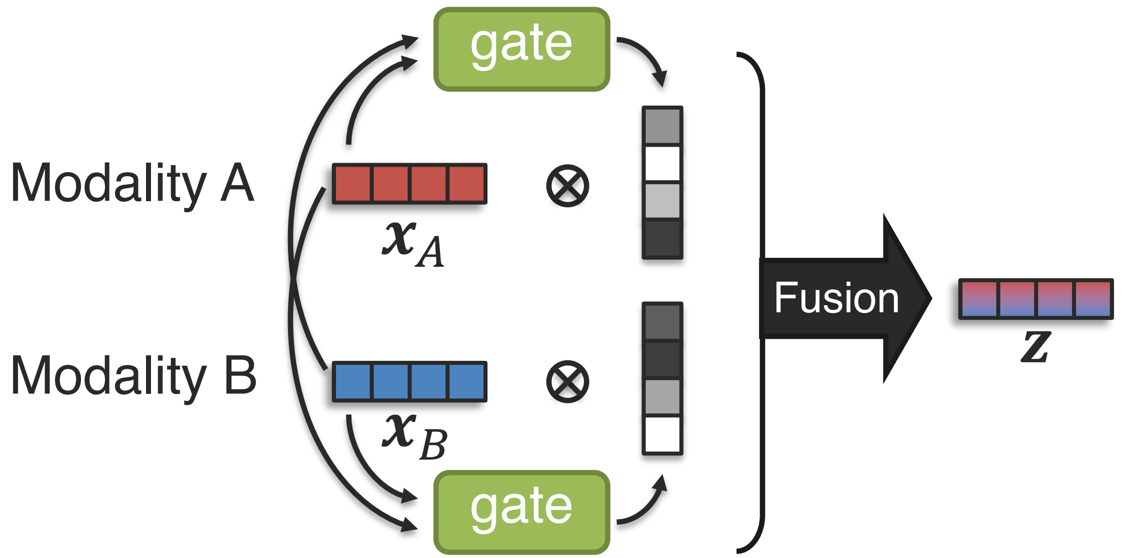
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Tensor Fusion



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

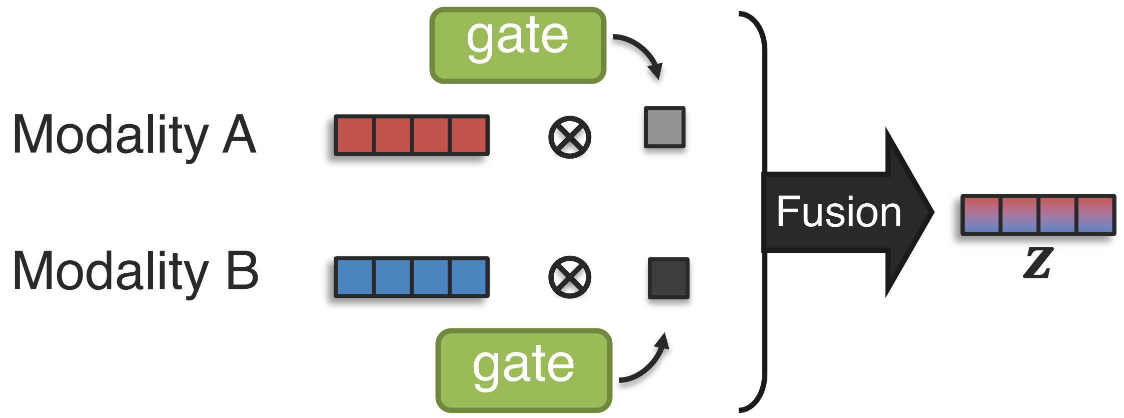
Gated Fusion



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



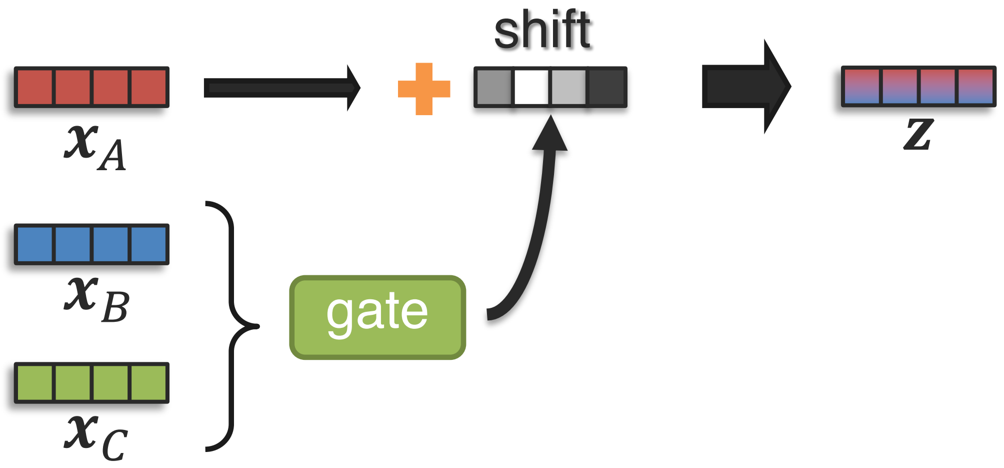
➡ Gating output can be one weight for the whole modality

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

Modality-Shifting Fusion

Primary modality

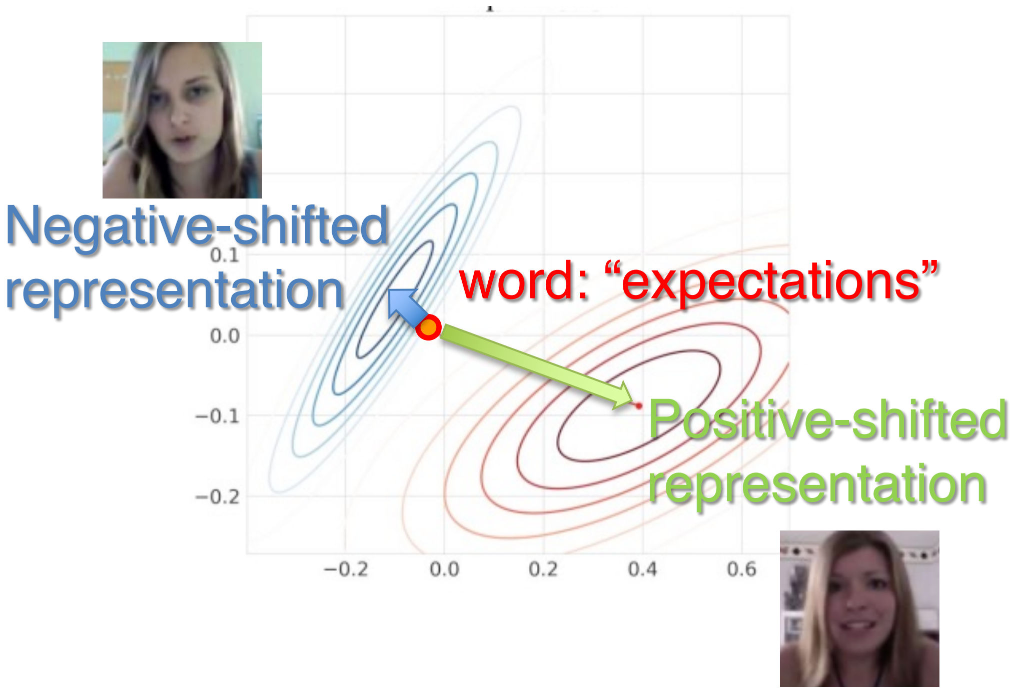
Secondary modalities



Example with language modality:

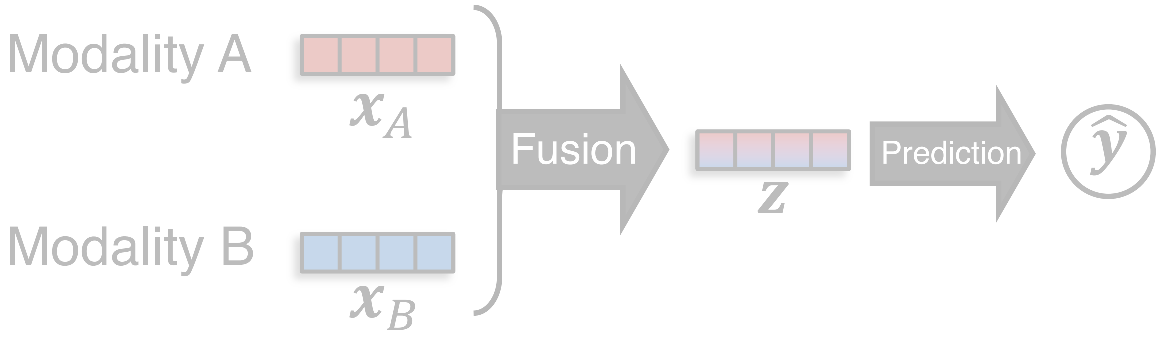
Primary modality: language

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]

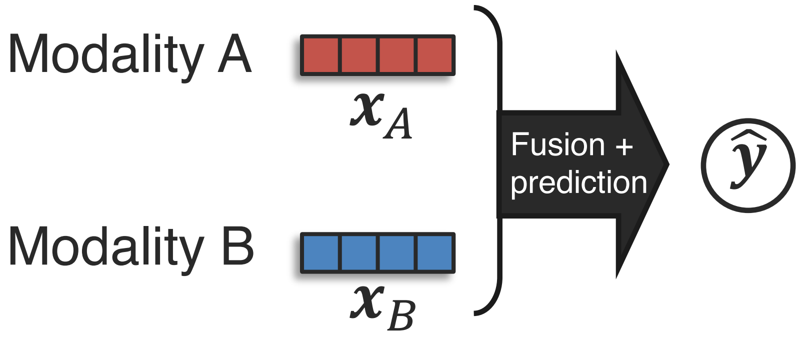
Nonlinear Fusion



Nonlinear fusion:

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

For any nonlinear model

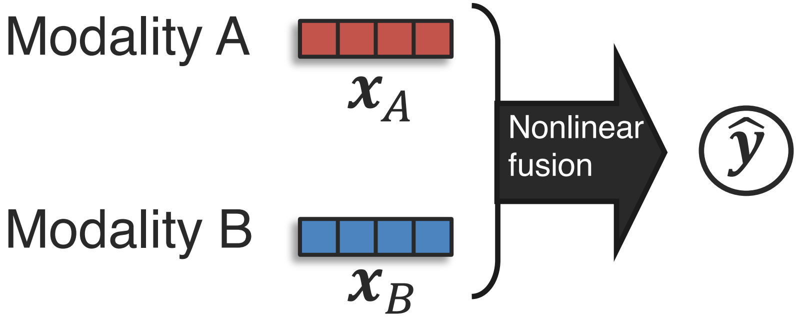


This could be seen as *early fusion*:

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

… but will our neural network learn the nonlinear interactions?

Measuring Non-Additive Interactions



Nonlinear fusion:

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

Projection?

Additive fusion:

$$\hat{y}' = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$$

Projection from nonlinear to additive (using EMAP):

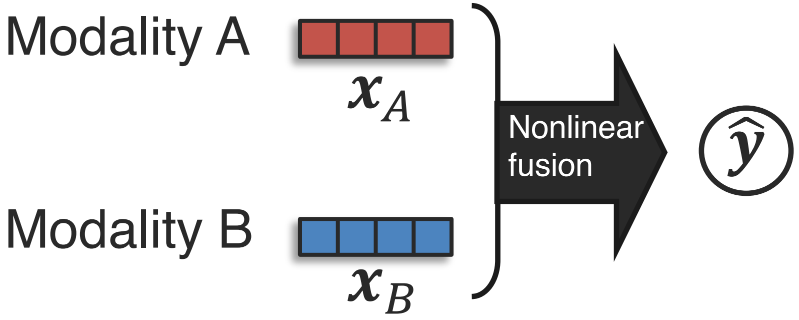
$$\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B) = \underbrace{\mathbb{E}_{\boldsymbol{x}_B}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_A(\boldsymbol{x}_A)} + \underbrace{\mathbb{E}_{\boldsymbol{x}_A}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_B(\boldsymbol{x}_B)}$$

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:

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

EMAP projection

Additive fusion:

$$\hat{y}' = \hat{f}_A(\boldsymbol{x}_A) + \hat{f}_B(\boldsymbol{x}_B) + \hat{\boldsymbol{\mu}}$$

		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

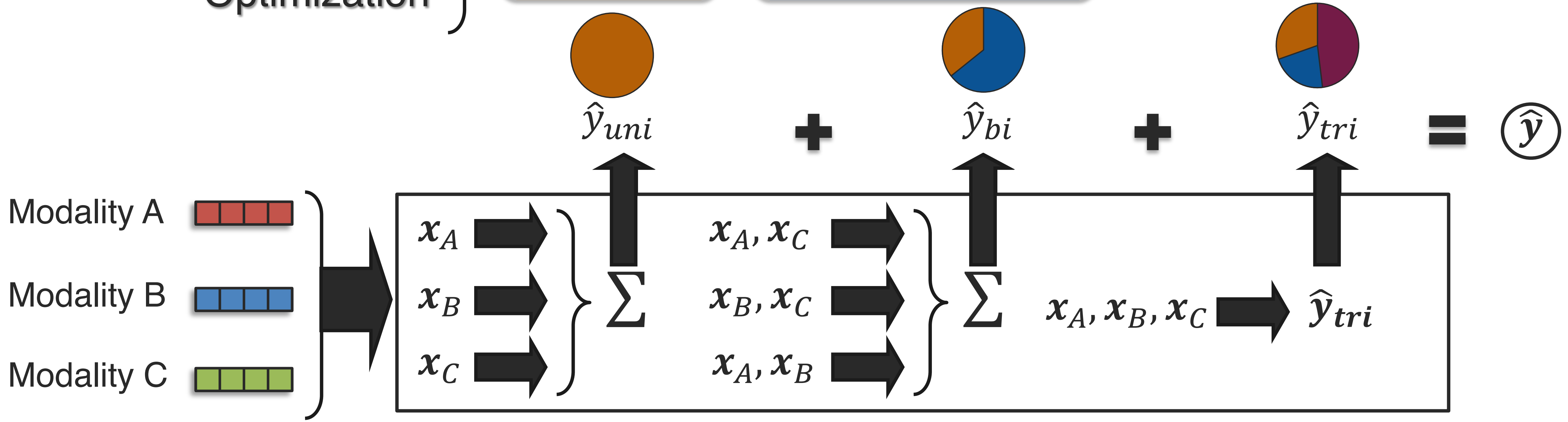
Idea: prioritize simpler interactions

Unimodal (additive) | Bimodal (non-additive) | Trimodal (non-additive)

residual | residual

Multimodal Residual Optimization

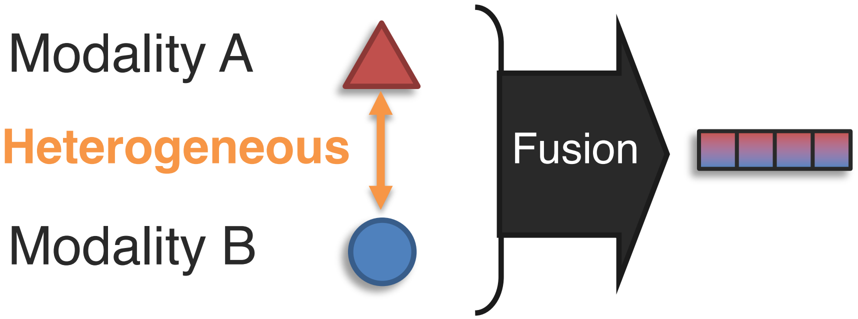
$$\mathcal{L}(y, \hat{y}_{uni}) + \mathcal{L}(y - \hat{y}_{uni}, \hat{y}_{bi}) + \mathcal{L}(y - \hat{y}_{uni} - \hat{y}_{bi}, \hat{y}_{tri})$$



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

Fusion with Heterogeneous Modalities

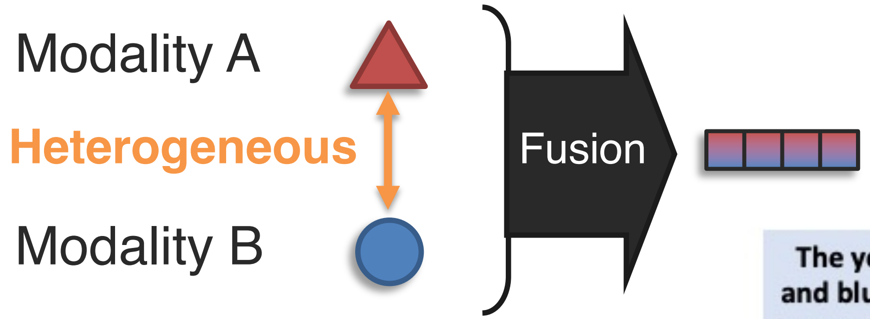
Example: From feature fusion to early fusion

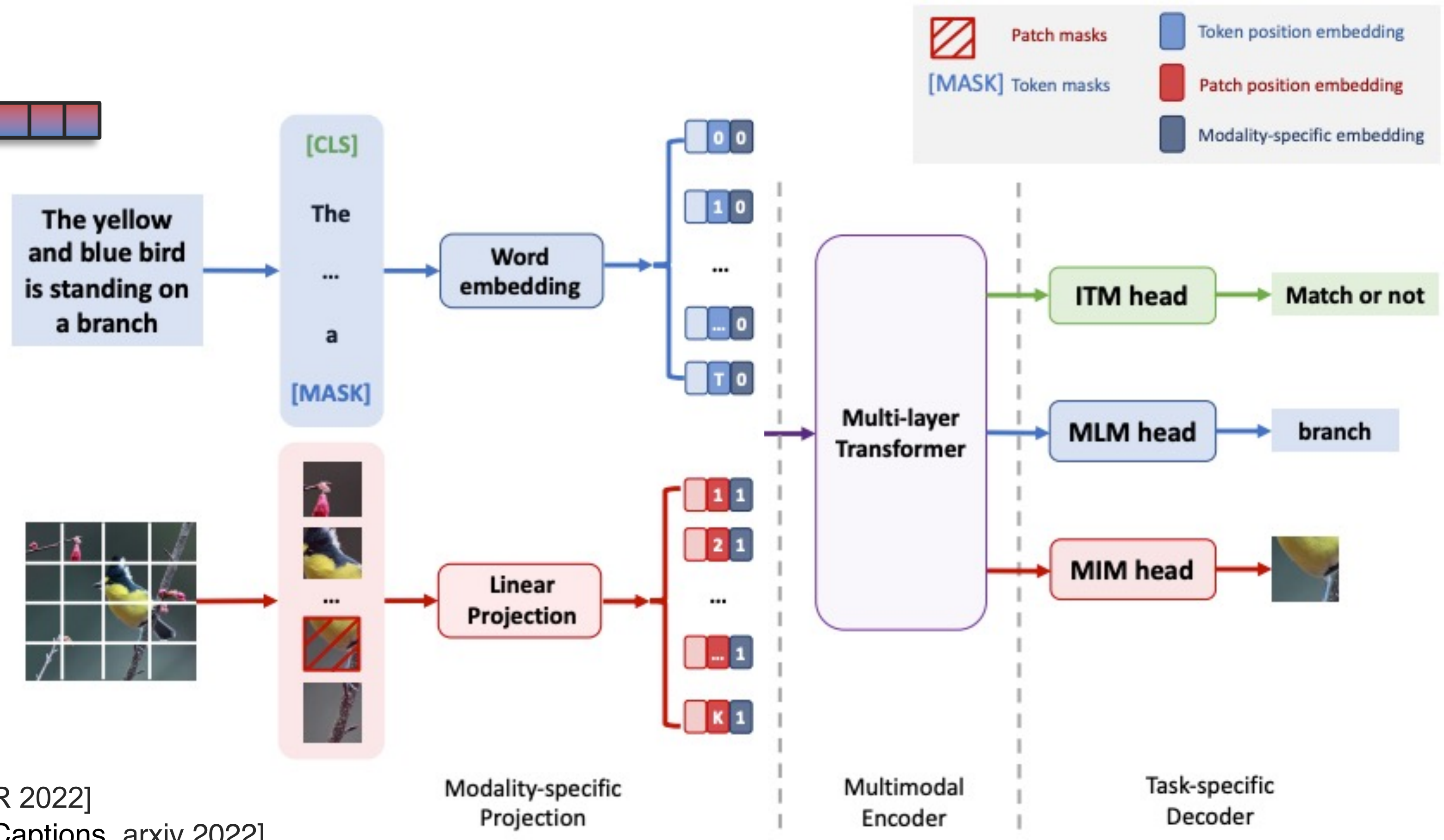


[Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR 2021]

Fusion with Heterogeneous Modalities

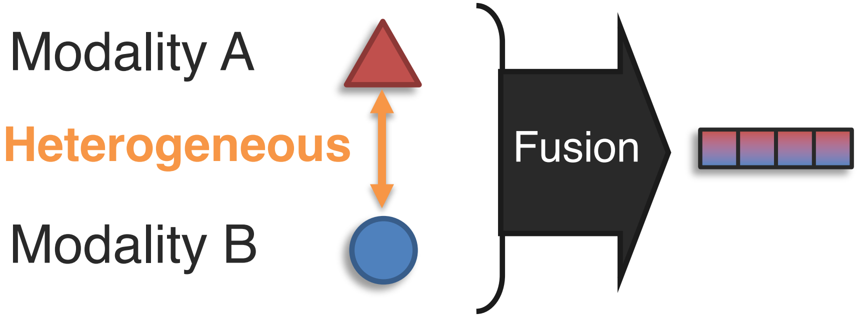
Example: From feature fusion to early fusion



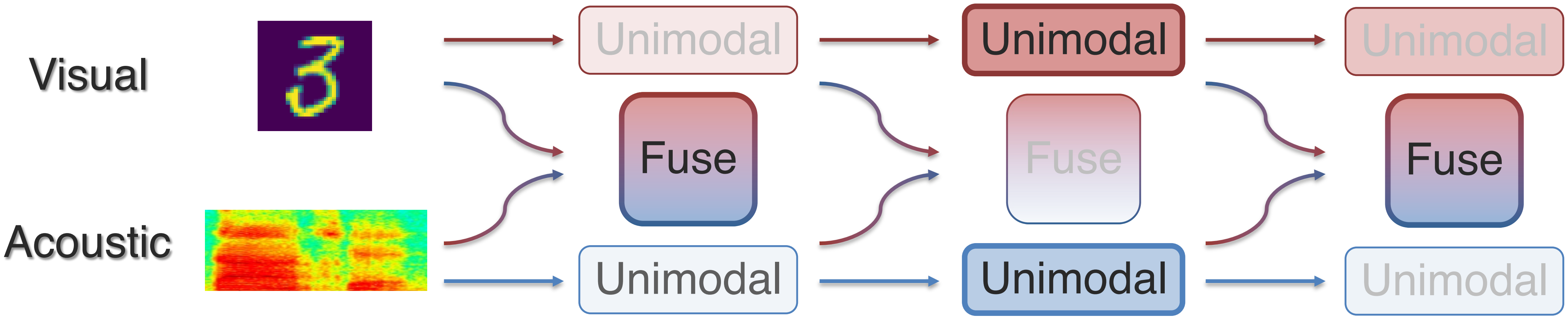


[Liang et al., High-modality Multimodal Transformer. TMLR 2022]
[Gui et al., Training Vision-Language Transformers from Captions. arxiv 2022]

Dynamic Early Fusion



Idea: Deciding when to fuse in early fusion



[Xue and Marculescu, Dynamic Multimodal Fusion, arxiv 2022]

Dynamic Early Fusion

Fusion fully learned from optimization and data

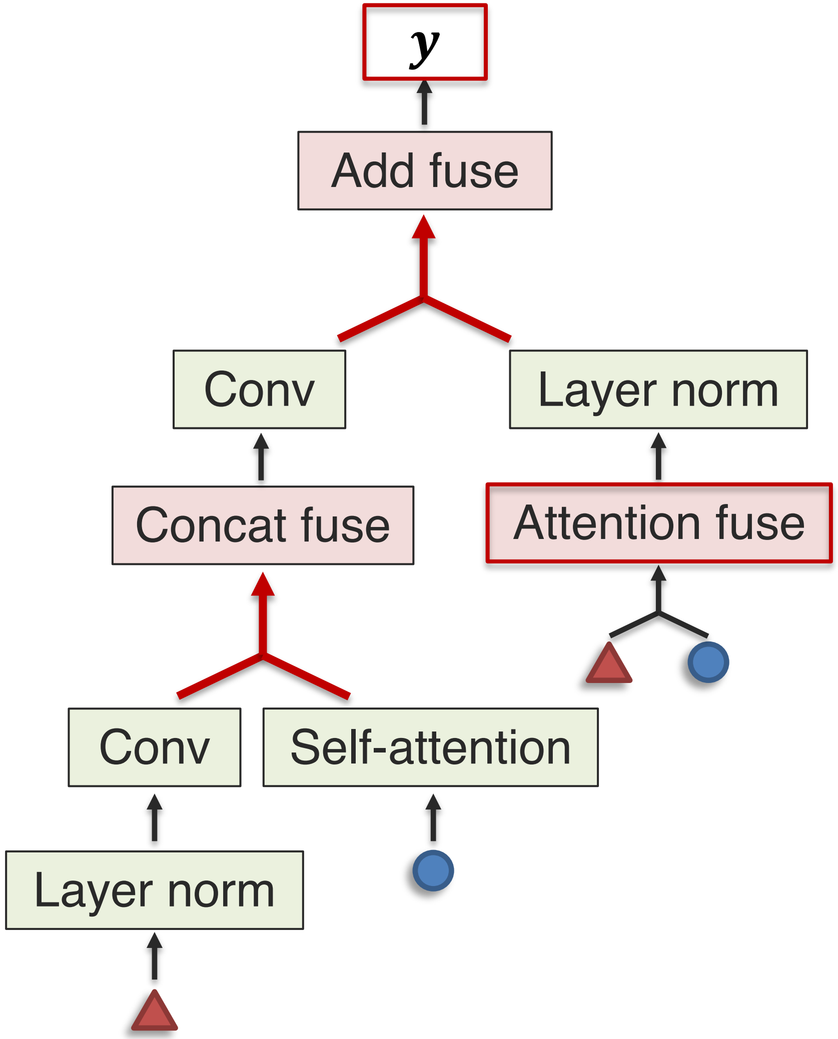
1. Define basic representation building blocks

ReLU | Layer norm | Conv | Self-attention

2. Define basic fusion building blocks

Concat fuse | Attention fuse | Add fuse

3. Automatically search for composition using neural architecture search

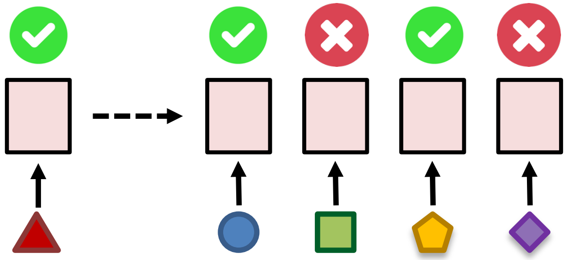


[Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records. AAAI 2021]
[Liu et al., DARTS: Differentiable Architecture Search. ICLR 2019]

Heterogeneity-aware Fusion

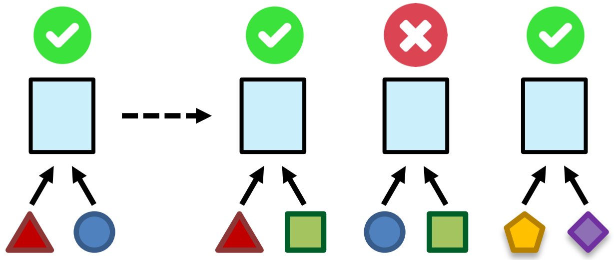
Information transfer, transfer learning perspective

1a. Estimate modality heterogeneity via transfer

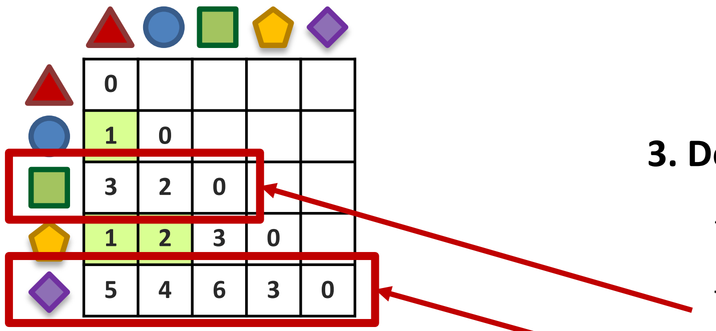


(Implicitly captures heterogeneity)

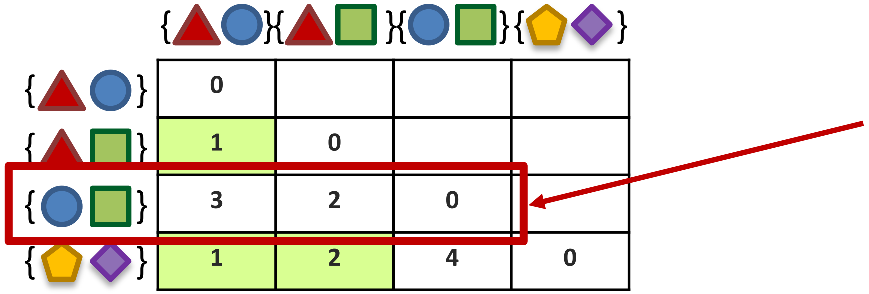
1b. Estimate interaction heterogeneity via transfer



2a. Compute modality heterogeneity matrix



2b. Compute interaction heterogeneity matrix



3. Determine parameter clustering

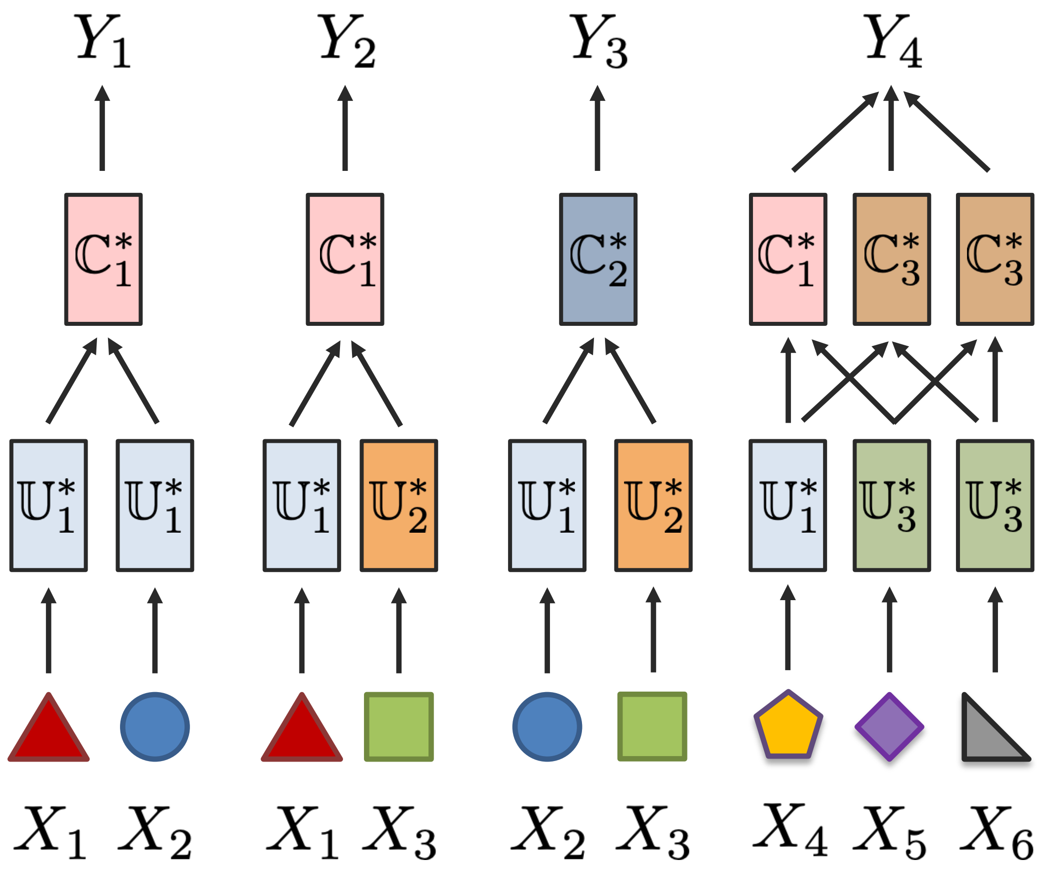
$$\mathbb{U}_1 = \{U_1, U_2, U_4\}$$
$$\mathbb{U}_2 = \{U_3\}$$
$$\mathbb{U}_3 = \{U_5\}$$

$$\mathbb{C}_1 = \{C_{12}, C_{13}, C_{45}\}$$
$$\mathbb{C}_2 = \{C_{23}\}$$

[Zamir et al., Taskonomy: Disentangling Task Transfer Learning. CVPR 2018]
[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

Heterogeneity-aware Fusion

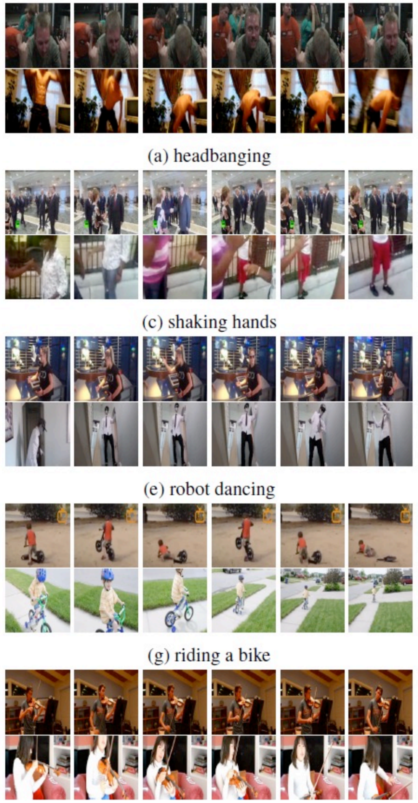
Information transfer, transfer learning perspective



[Liang et al., HighMMT: Quantifying Modality & Interaction Heterogeneity for High-Modality Learning. TMLR 2022]

Improving Optimization

Kinetics dataset



Adding more modalities should always help?

Modalities: **RGB** (video clips)

A (Audio features)

OF (optical flow - motion)

Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6

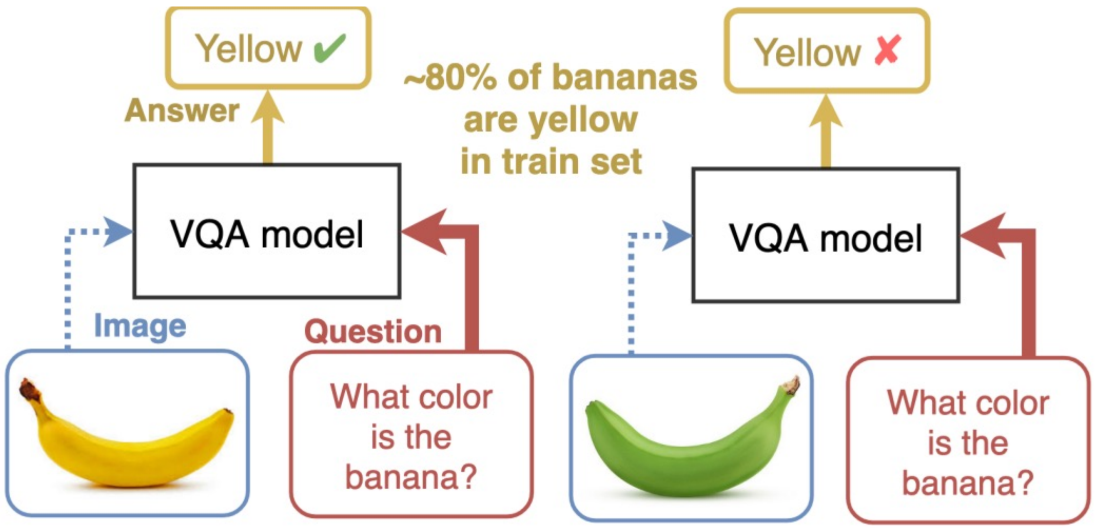
But sometimes multimodal doesn't help! **Why?**

[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]

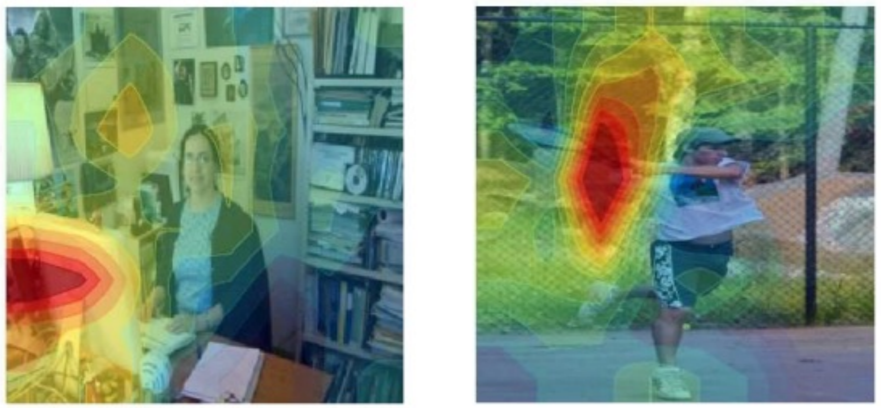
Improving Optimization

Information heterogeneity and unimodal biases

Finding: VQA models answer the question without looking at the image



Finding: Image captioning models capture spurious correlations between gender and generated actions.



Baseline:
*A **man** sitting at a desk with a laptop computer.*

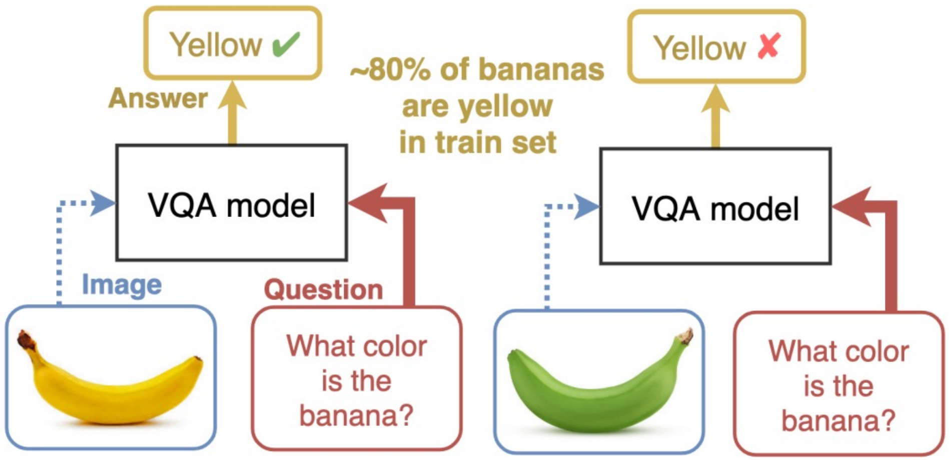
Baseline:
*A **man** holding a tennis racquet on a tennis court.*

[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017]
[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

Improving Optimization

Information heterogeneity and modality collapse

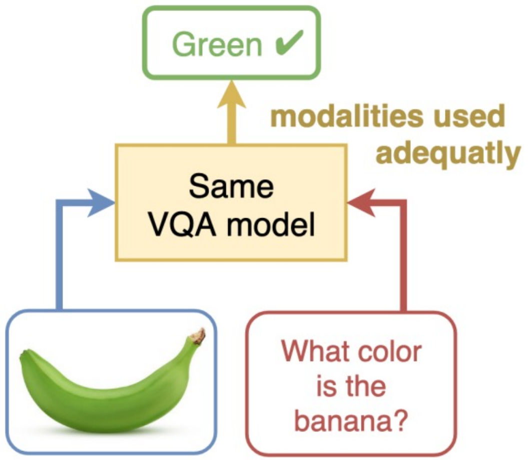
VQA models answer the question without looking at the image



Balancing modalities

Balancing training

Not the case when trained with RUBi



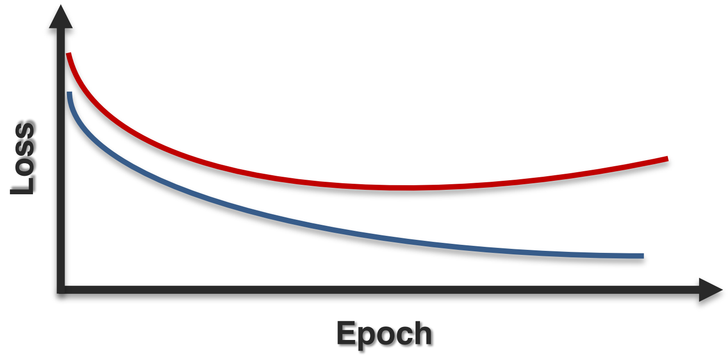
[Javaloy et al., Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. ICML 2022]
[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017]

Improving Optimization

Relevance heterogeneity

2 explanations for drop in performance:

1. Multimodal networks are more prone to overfitting due to **increased complexity**
2. Different modalities overfit and generalize at **different rates**



Key idea 1: compute overfitting-to-generalization ratio (OGR)

Gap between training and valid loss

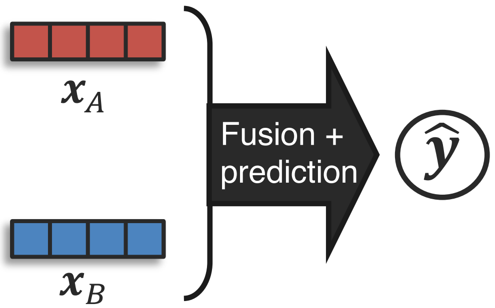
OGR wrt each modality tells us how much to train that modality

[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]

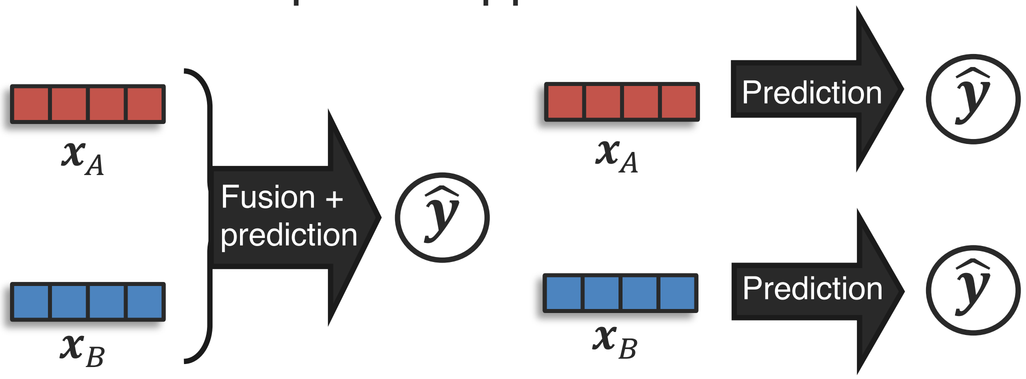
Improving Optimization

Relevance heterogeneity

Conventional approach



Proposed approach



Key idea 2: Simultaneously train unimodal networks to estimate OGR wrt each modality

Reweight multimodal loss using unimodal OGR values

Allows to better balance generalization & overfitting rate of different modalities

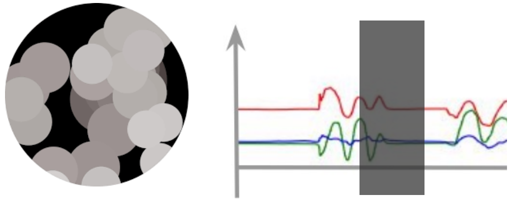
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]

Improving Robustness

Heterogeneity in noise

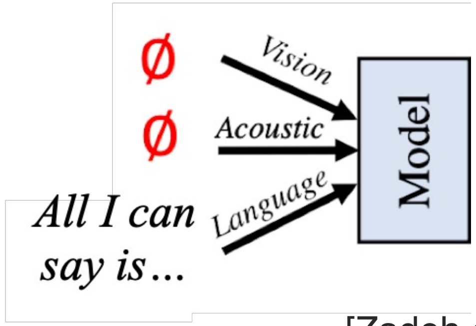
Modality-specific robustness

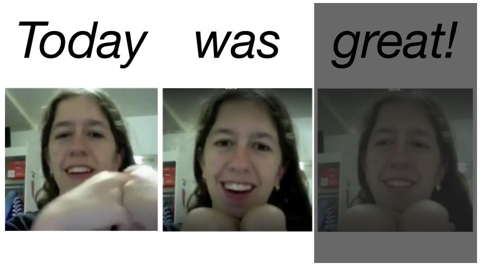
noise ⟶ nosie



[Belinkov & Bisk, 2018; Subramaniam et al., 2009; Boyat & Joshi, 2015]

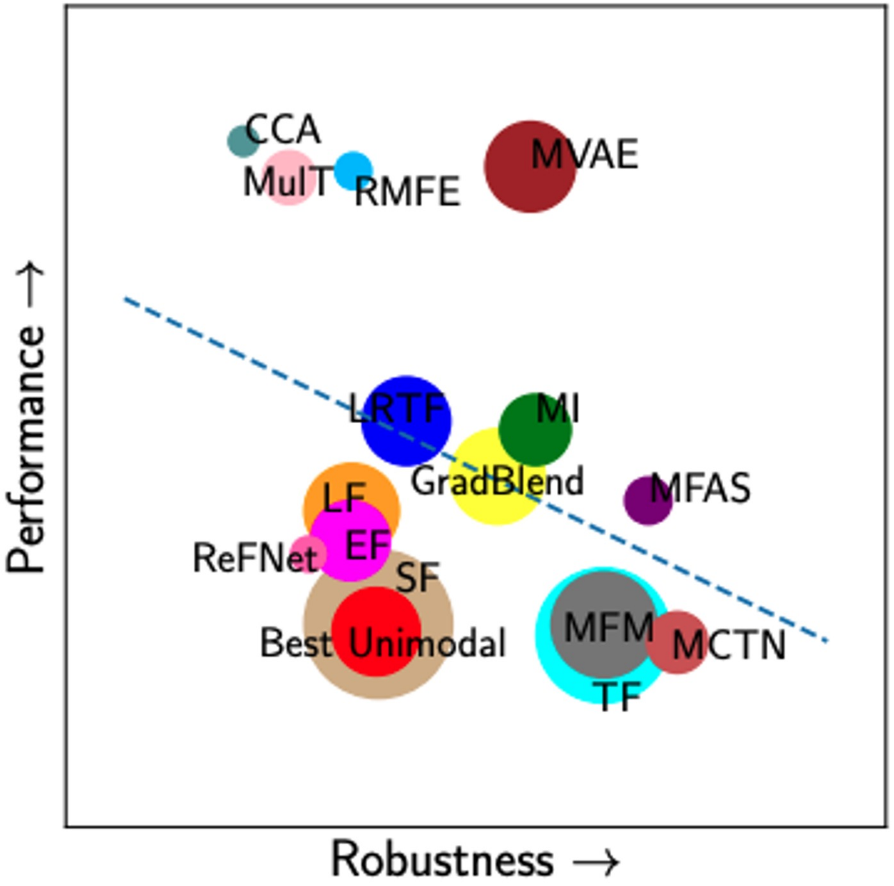
Multimodal robustness





[Zadeh et al., 2020]

Tradeoffs between performance and robustness

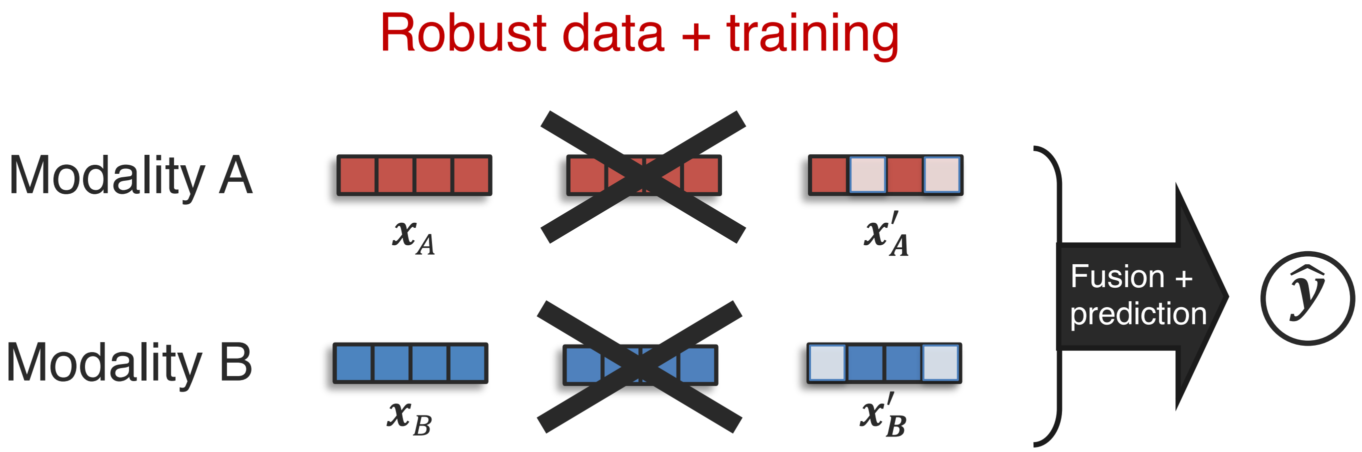


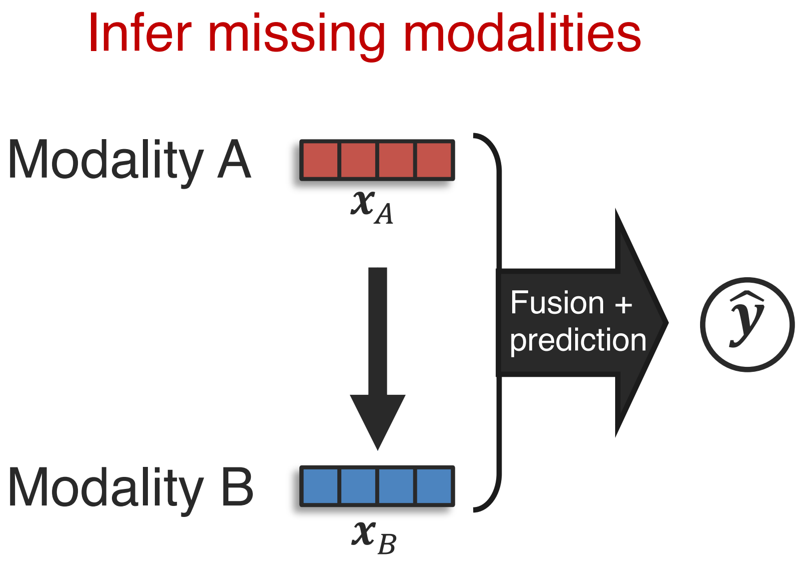
rate of accuracy drops

[Liang et al., MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. NeurIPS 2021]

Improving Robustness

Several approaches towards more robust models

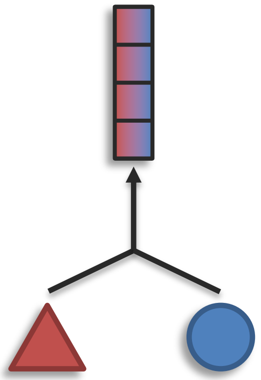




Translation model
Joint probabilistic model

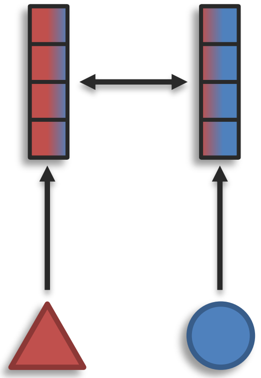
[Ngiam et al., Multimodal Deep Learning. ICML 2011]
[Srivastava and Salakhutdinov, Multimodal Learning with Deep Boltzmann Machines. JMLR 2014]
[Tran et al., Missing Modalities Imputation via Cascaded Residual Autoencoder. CVPR 2017]
[Pham et al., Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities. AAAI 2019

Sub-Challenge 1a: Representation Fusion

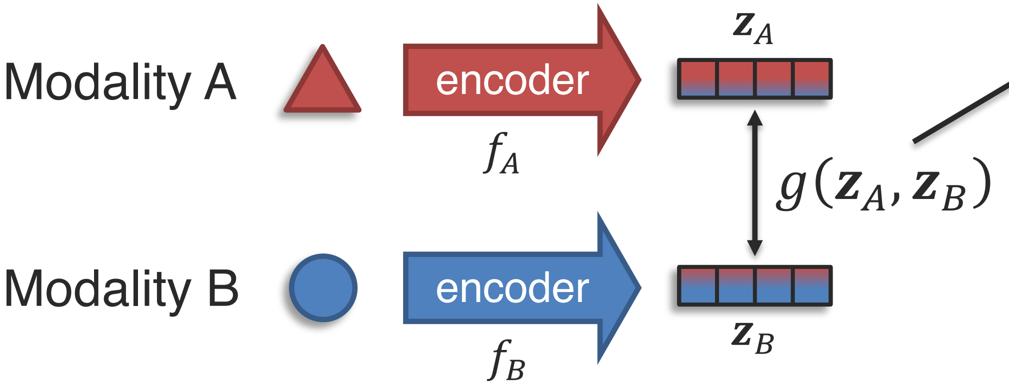


Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Sub-Challenge: Representation Coordination



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



2. Coordination function captures **interconnections**

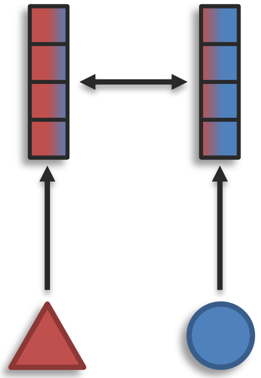
1. Specialized encoders capture **heterogeneity**

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}

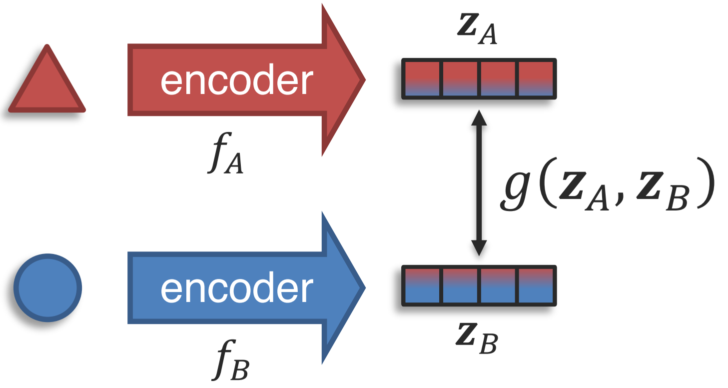
Sub-Challenge: Representation Coordination



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

Modality A

Modality B



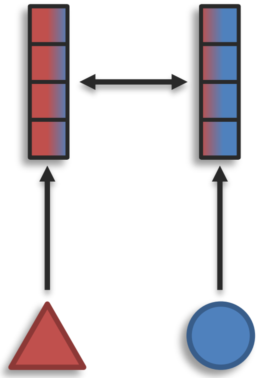
Learning with coordination function:

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

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

1. Cosine similarity: $g(\mathbf{z}_A, \mathbf{z}_B) = \dfrac{\mathbf{z}_A \cdot \mathbf{z}_B}{\|\mathbf{z}_A\| \|\mathbf{z}_B\|}$

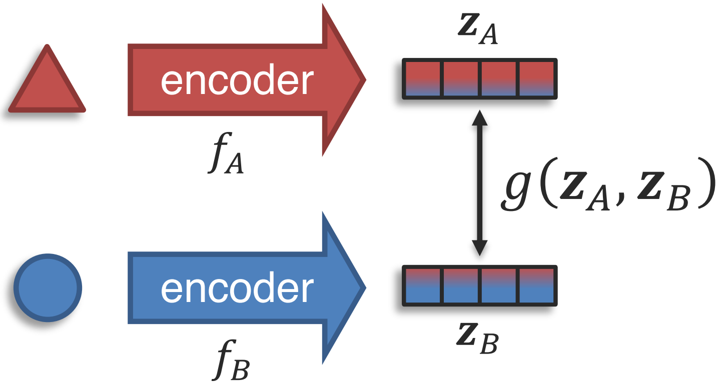
Sub-Challenge: Representation Coordination



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

Modality A

Modality B



Learning with coordination function:

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

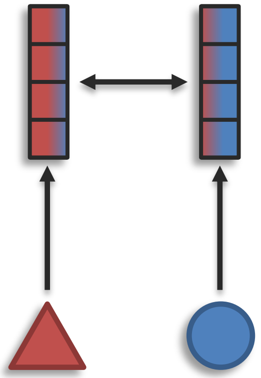
with model parameters θ_g, θ_{f_A} and θ_{f_B}

② Kernel similarity functions:

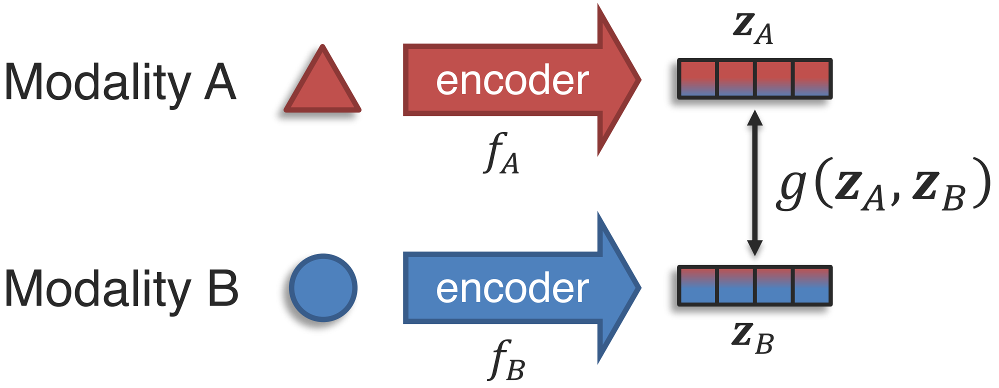
$$g(\mathbf{z}_A, \mathbf{z}_B) = k(\mathbf{z}_A, \mathbf{z}_B) \left\{ \right.$$

- Linear
- Polynomial
- Exponential
- RBF

Sub-Challenge: Representation Coordination



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



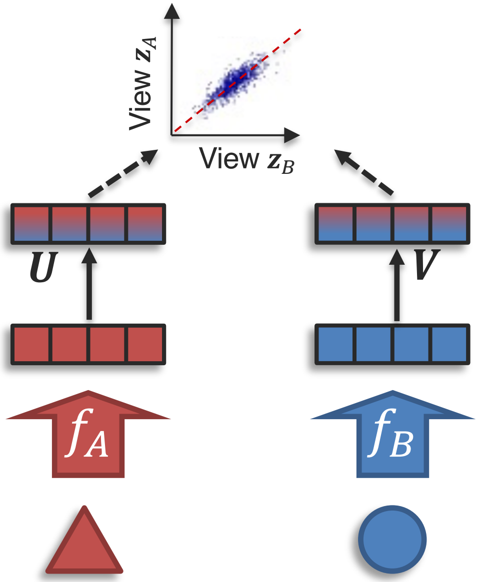
Learning with coordination function:

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

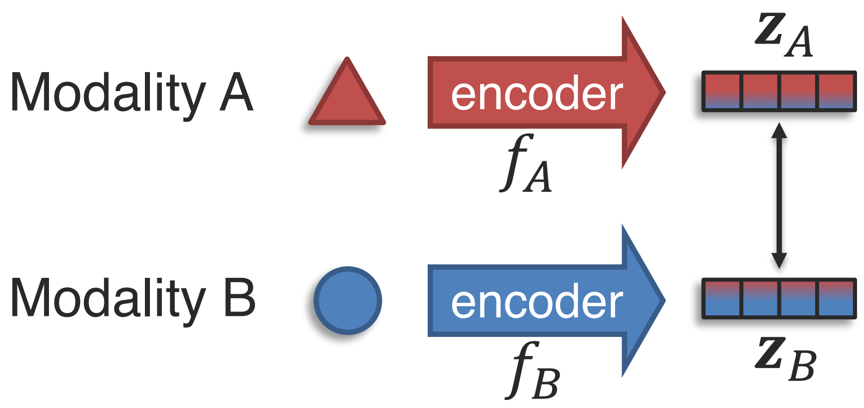
with model parameters θ_g, θ_{f_A} and θ_{f_B}

③ Canonical Correlation Analysis:

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

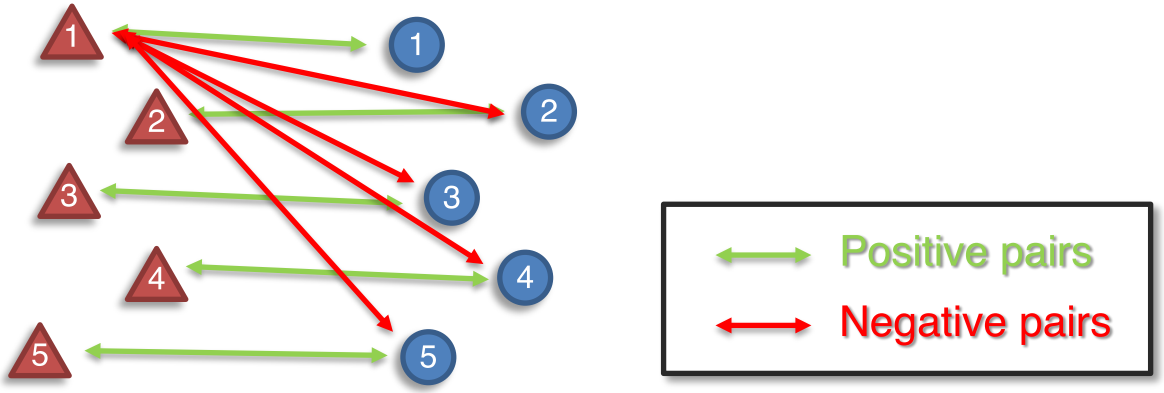


Coordination with Contrastive Learning



Paired data: 

(e.g., images and text descriptions)



Contrastive loss:

brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

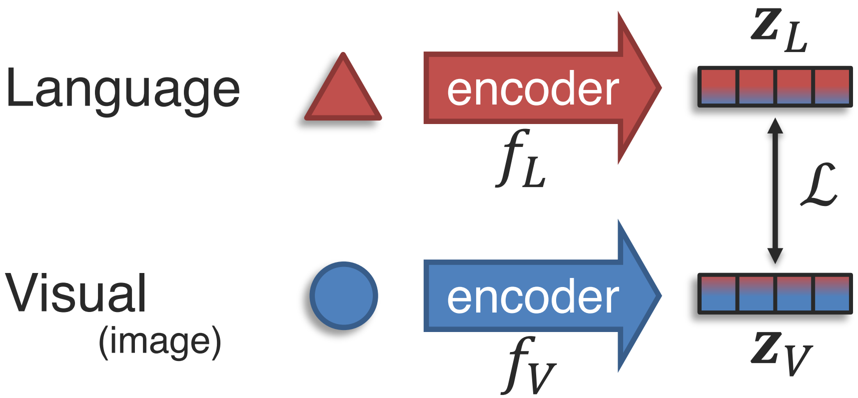
$$\max\{0, \alpha + \boxed{sim(\boldsymbol{z}_A, \boldsymbol{z}_B^+)} - \boxed{sim(\boldsymbol{z}_A, \boldsymbol{z}_B^-)}\}$$

positive pairs

negative pair

Similarity functions are often cosine similarity

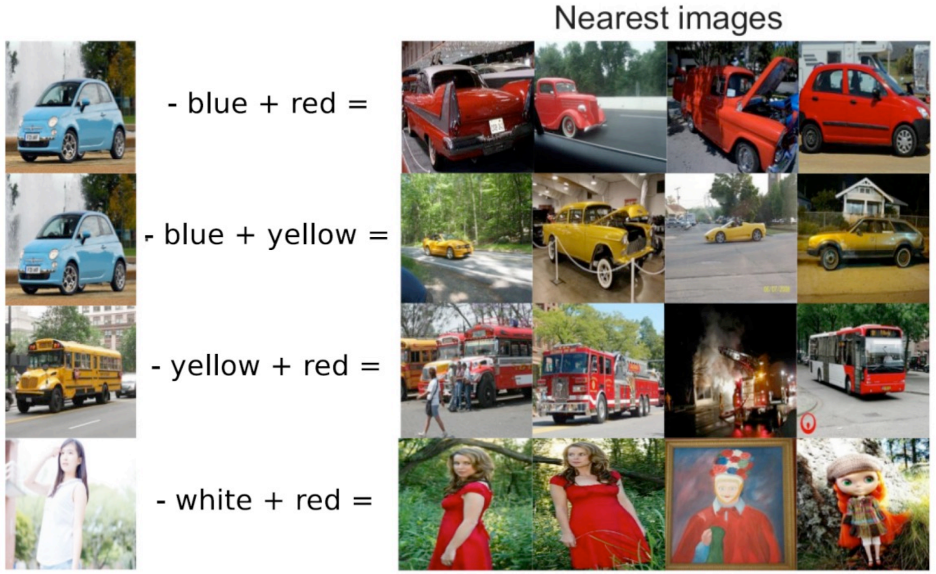
Example – Visual-Semantic Embeddings

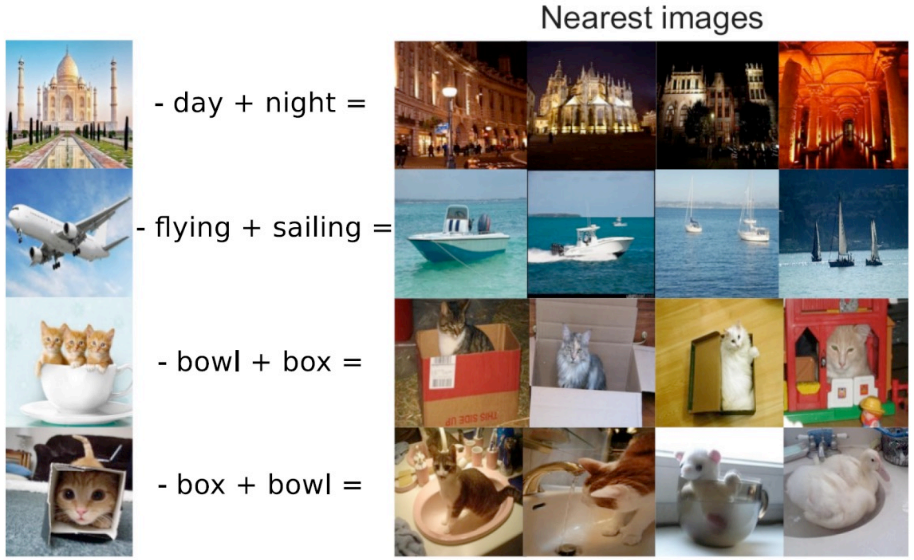


Two contrastive loss terms:

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

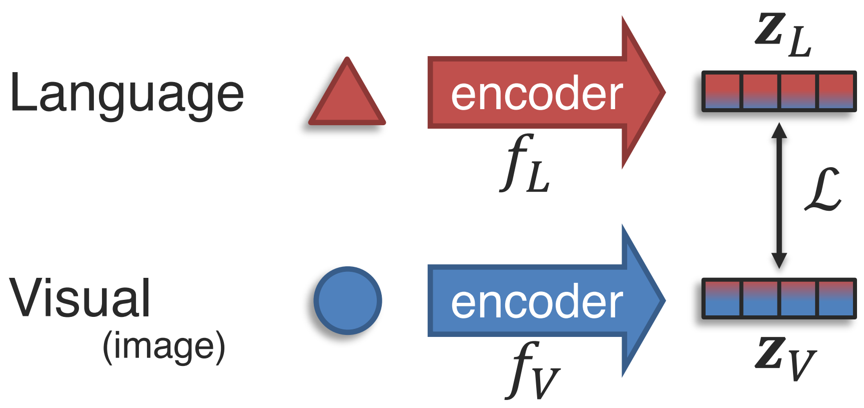
$$+ \max\{0, \alpha + sim(\boldsymbol{z}_V, \boldsymbol{z}_L^{+}) - sim(\boldsymbol{z}_V, \boldsymbol{z}_L^{-})\}$$





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

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



Popular contrastive loss: InfoNCE

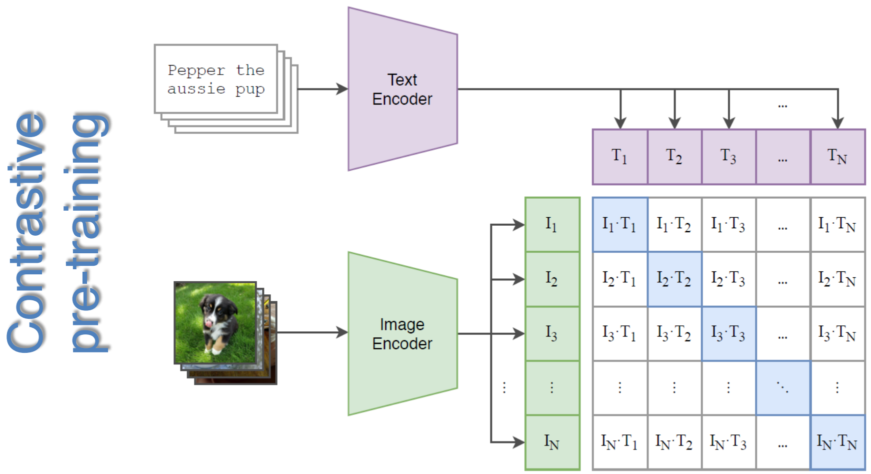
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{\boxed{sim(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}}{\boxed{\sum_{j=1}^{N} sim(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}}$$

positive pairs

Similarity function can be cosine similarity

negative pairs and positive pairs

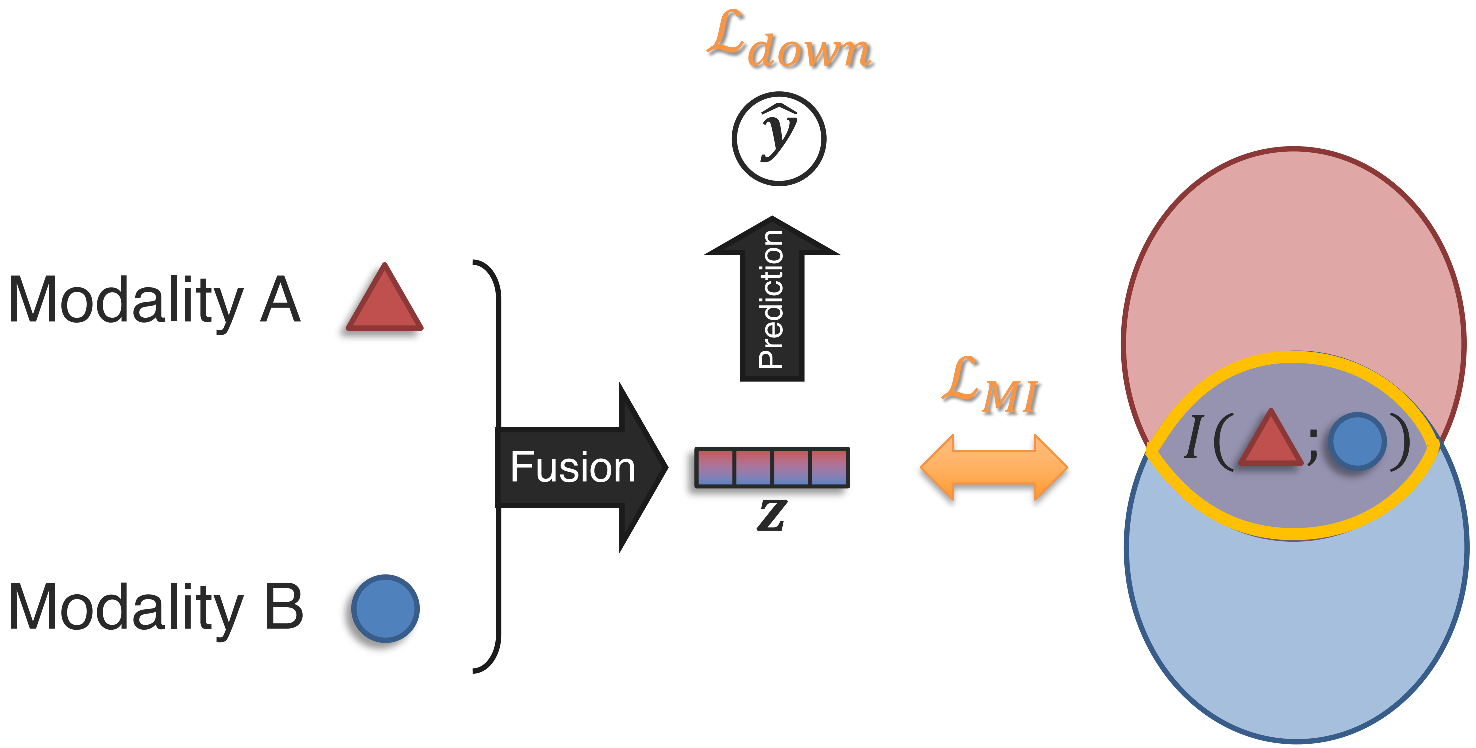
Positive and negative pairs:



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

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

Multimodal Fusion with Mutual Information

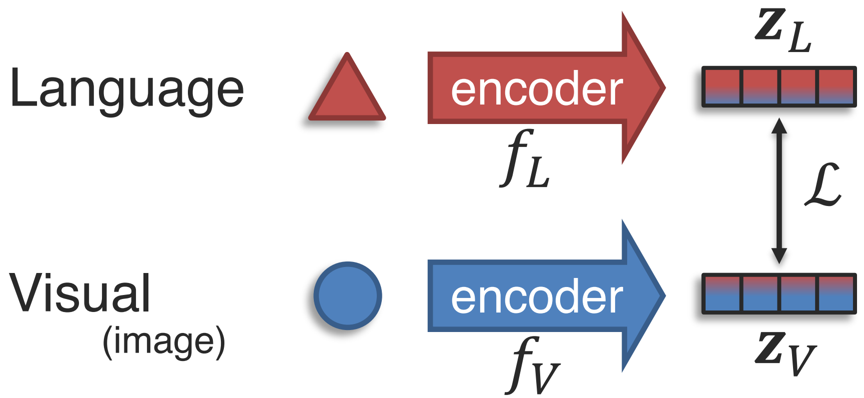


Assumption?

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

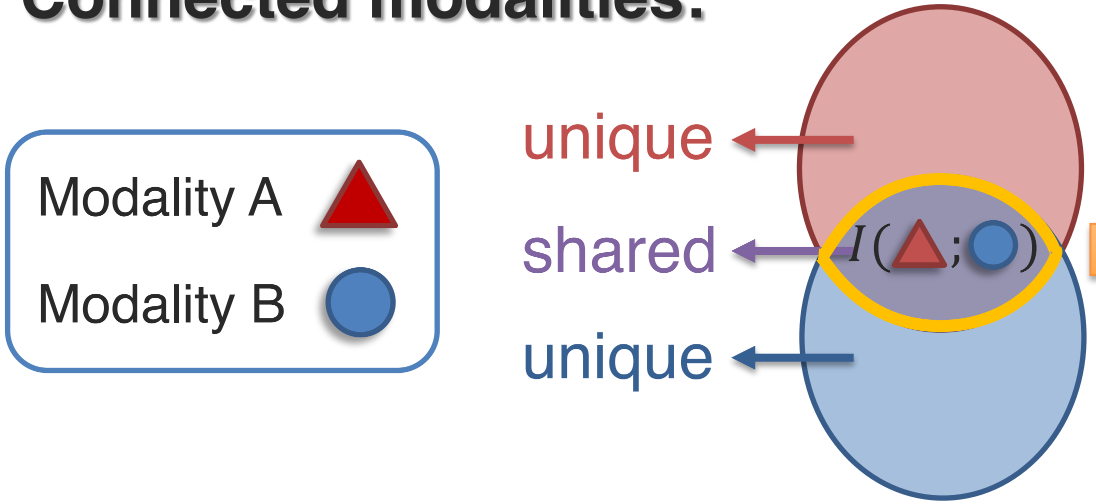


Popular contrastive loss: InfoNCE

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

CLIP focuses on shared connections

Connected modalities:

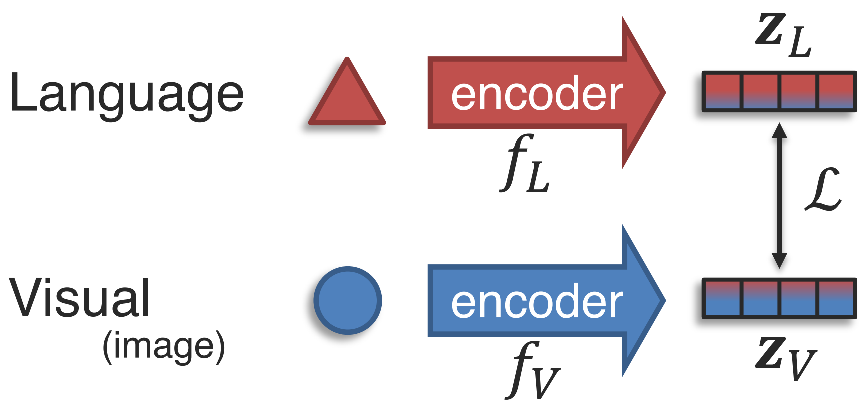


Mutual information $I(X;Y)$

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

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

Contrastive Learning and Mutual Information



InfoNCE:

critic function

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

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

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

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

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

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

InfoNCE/CL:

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

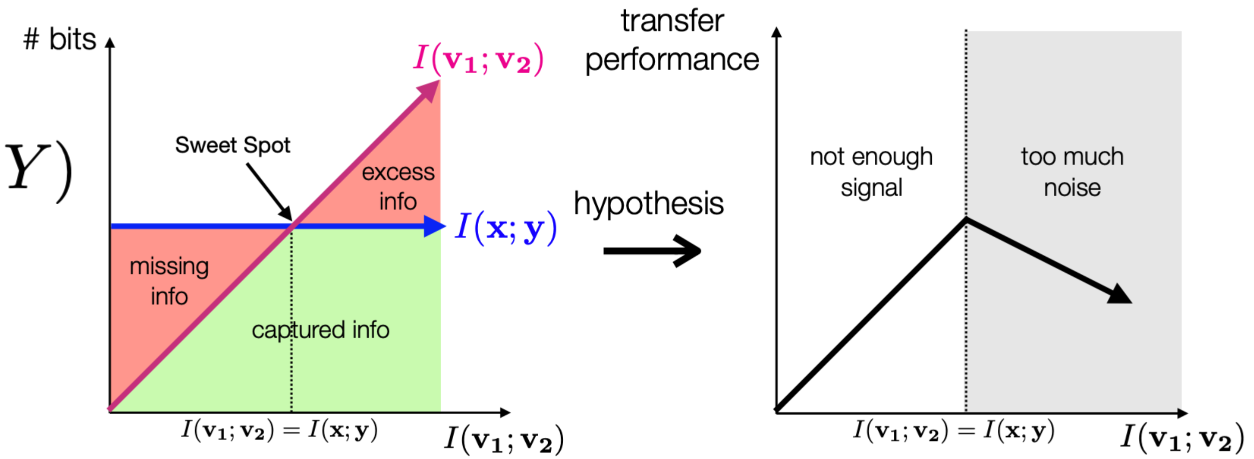
[Oord et al., Representation Learning with Contrastive Predictive Coding. 2018]

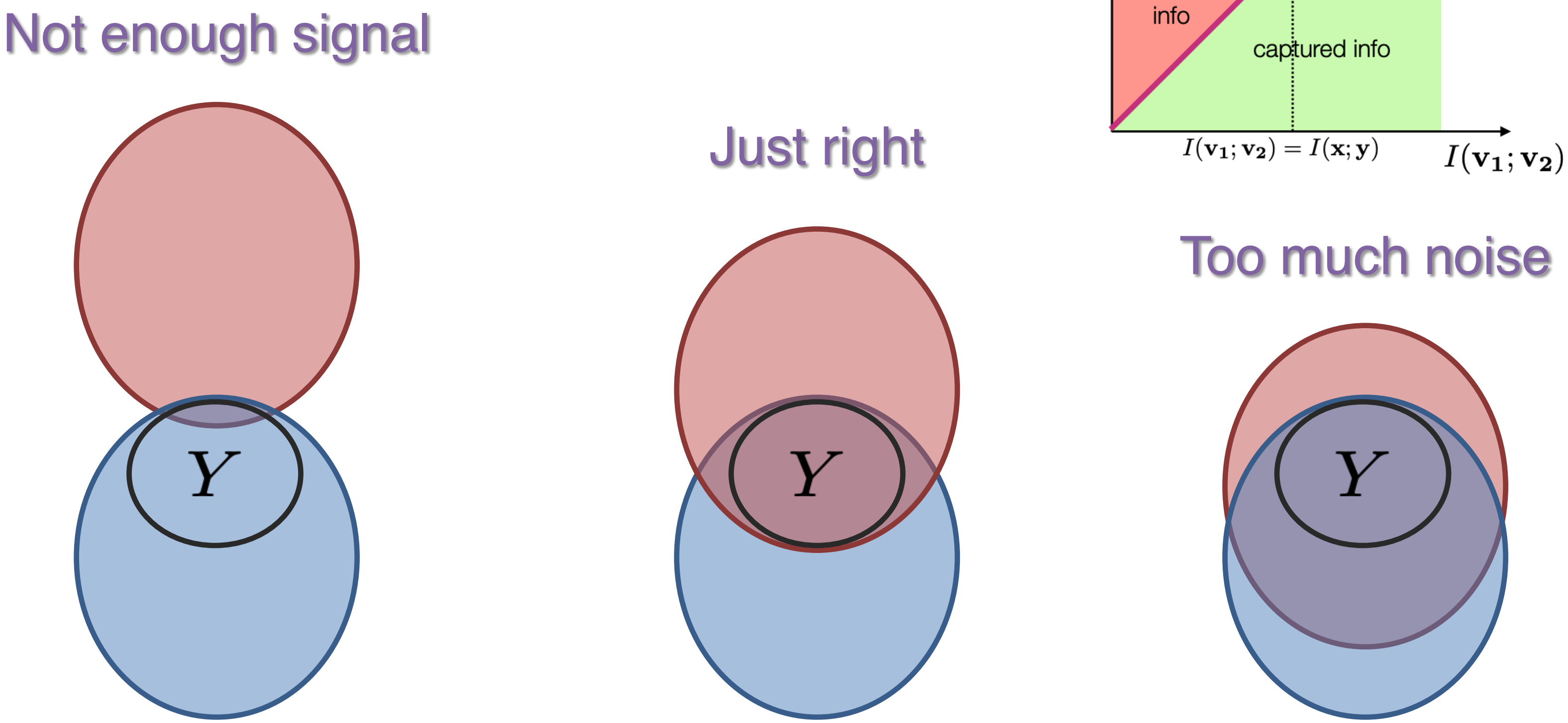
Multiview Redundancy and Contrastive Learning

Open challenges

How much information should be shared?

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

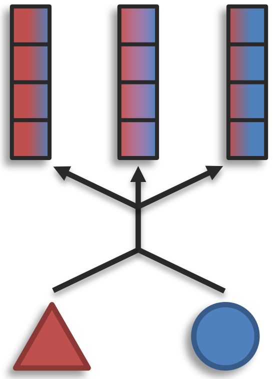




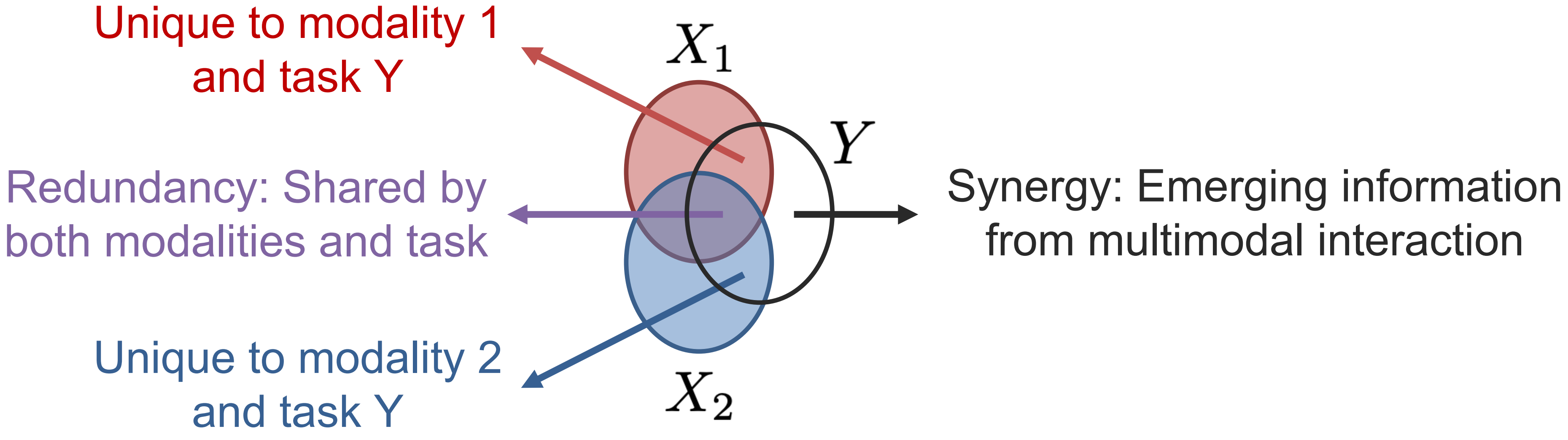
Multi-view redundancy may not hold for multimodal problems!

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

Sub-Challenge 1c: Representation Fission



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

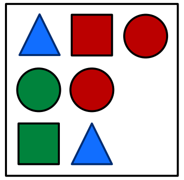


Quantifying Interactions

These interactions can be estimated efficiently:

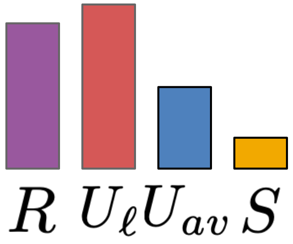
Sheldon :
Its just a *privilege* to watch your mind at work.

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



Is there a red shape above a circle?

Sentiment

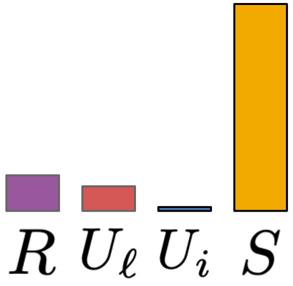


Language/Agreement

Sarcasm

Multimodal Transformer

VQA



Multiplicative/Transformer

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

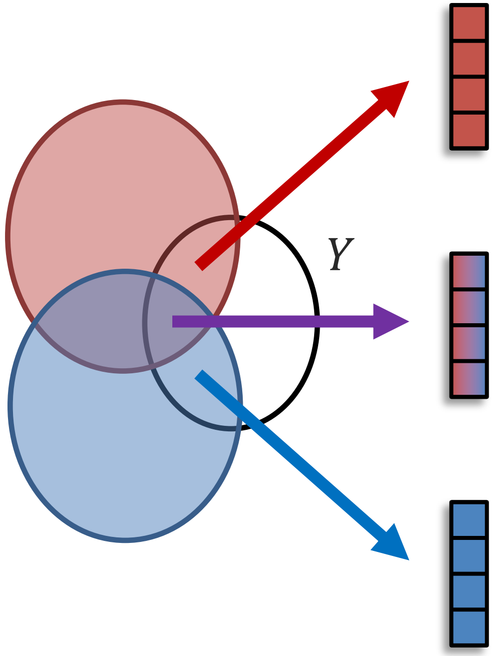
Can also be used to choose most appropriate models – can they be used to better train/design new models?

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

Factorized Contrastive Learning

Modeling task-relevant unique information

Can you please pass the cow?



2 Maximize task-relevant **unique** information

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

1 Maximize task-relevant **shared** information

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

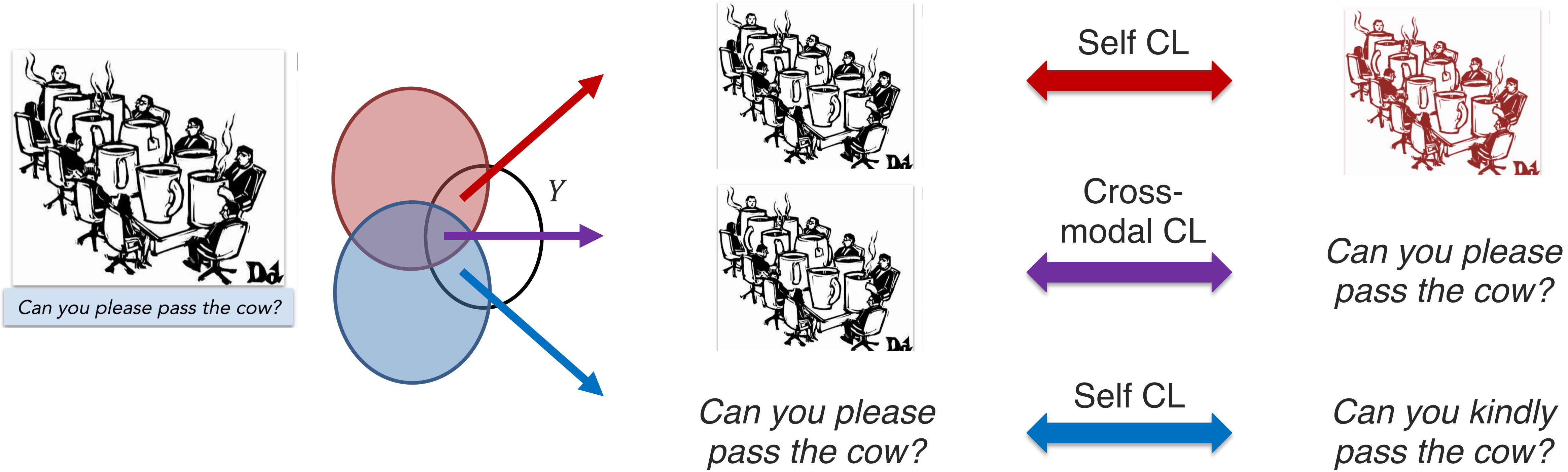
3 Maximize task-relevant **unique** information

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

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

Factorized Contrastive Learning

Modeling task-relevant unique information



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

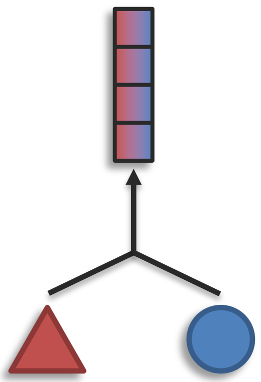
[Liang et al., Factorized Contrastive Learning: Going Beyond Multi-view Redundancy, NeurIPS 2023]

Challenge 1: Representation

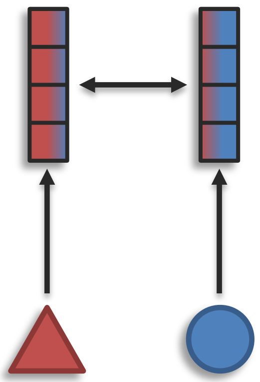
Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

Sub-challenges:

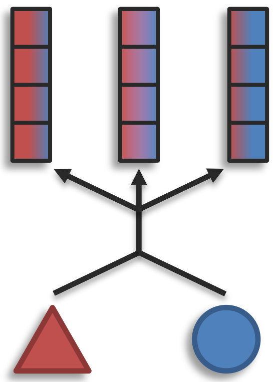
Fusion



Coordination



Fission



What is Multimodal? → Why is it hard? → What is next?

What is Multimodal?

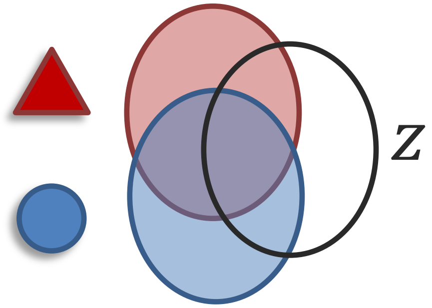
Heterogeneous

+

Connected

+

Interacting



Why is it hard?

Representation

Alignment

Reasoning

Generation

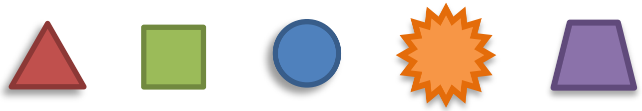
Transference

Quantification

What is next?

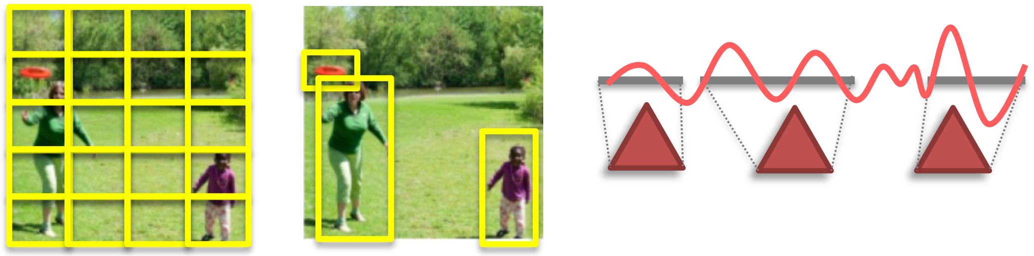
Future Direction: Heterogeneity

Homogeneity vs Heterogeneity



Examples:

Arbitrary tokenization



Beyond differentiable interactions

Causal, logical, brain-inspired

Theoretical studies

Future Direction: High-modality

MultiBench

https://github.com/pliang279/MultiBench

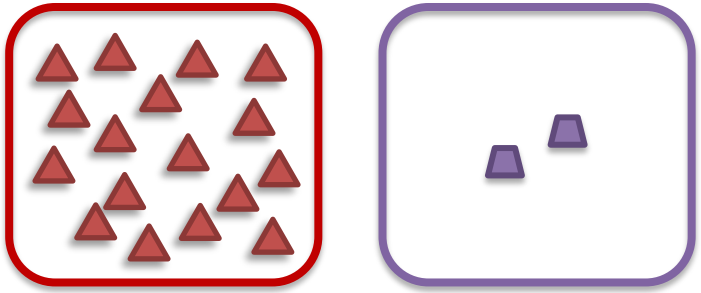
Few modalities → High-modality

Language Vision Audio Graphs Control LIDAR Sensors Set Table Financial Medical

Examples:

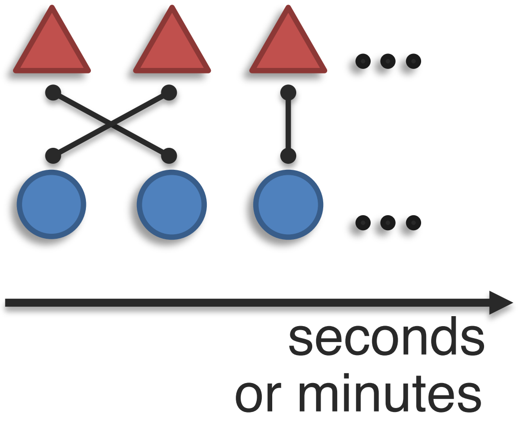
Non-parallel learning

Limited resources



Future Direction: Long-term

Short-term



Long-term

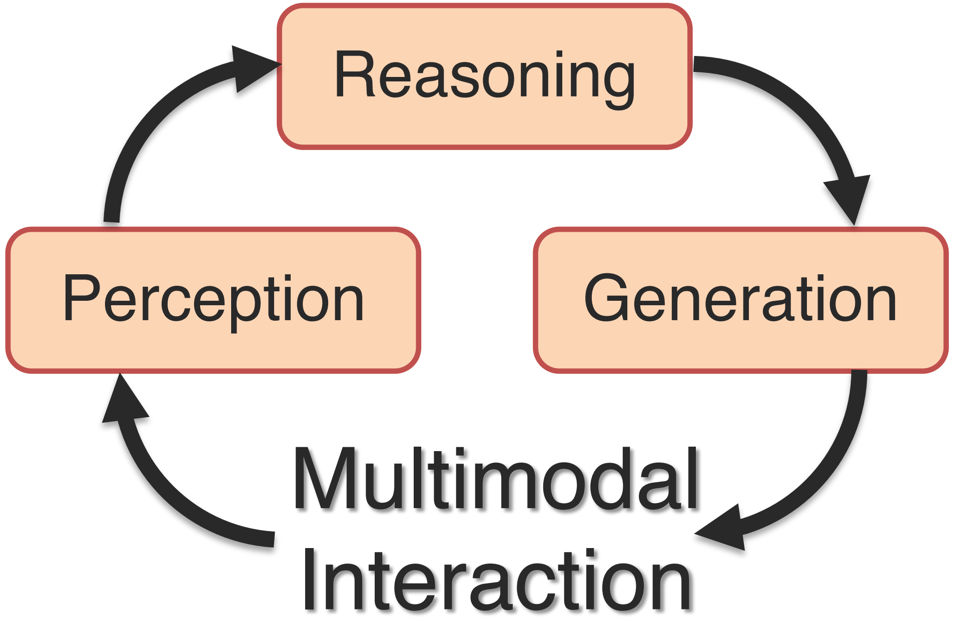
Examples:

Compositionality **Memory** **Personalization**

Future Direction: Interaction

Social-IQ

https://www.thesocialiq.com/



Social Intelligence

Examples:

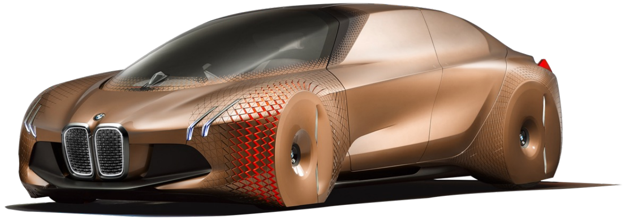
Multi-Party **Causality** **Ethical**

MultiViz

https://github.com/pliang279/MultiViz

Future Direction: Real-world

Healthcare Decision Support



Intelligent Interfaces and Vehicles

Online Learning and Education

Examples:

Robustness **Fairness** **Generalization** **Interpretation**

What is Multimodal? → Why is it hard? → What is next?

What is Multimodal?

Heterogeneous

+

Connected

+

Interacting

Why is it hard?

Representation

Alignment

Reasoning

Generation

Transference

Quantification

What is next?

High-modality

Heterogeneity

Long-term

Interaction

Real-world

https://cmu-multicomp-lab.github.io/mmml-course/fall2023/

Liang, Zadeh, and Morency. Foundations and Trends on Multimodal Machine Learning. 2022

https://www.cs.cmu.edu/~pliang/

pliang@cs.cmu.edu

@pliang279