Foundations & Recent Trends in Multimodal Machine Learning

Paul Pu Liang
Machine Learning Department
Carnegie Mellon University

Carnegie Mellon University

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

MACHINE LEARNING DEPARTMENT

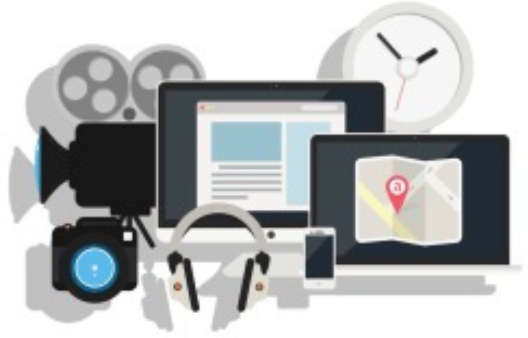
Real-world Artificial Intelligence

Digital intelligence

Multimedia

Image/video description

[Rui et al., 1999; Huang 2004]



Physical intelligence

Embodied AI, autonomous driving

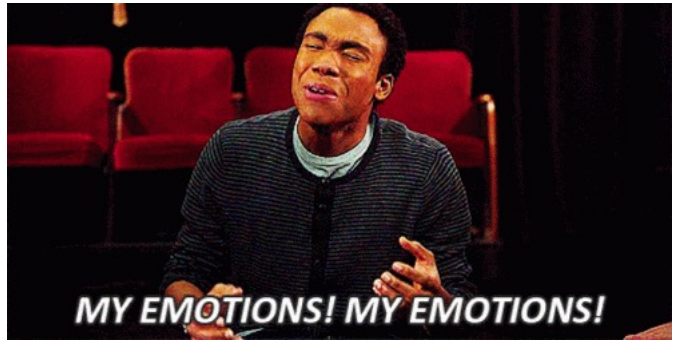
[Xu et al., 2017; Szot et al., 2021]

Social intelligence

Affective computing

Human-AI interaction

[Picard 1997; Jaimes & Sebe 2007]



Multimodal Artificial Intelligence

Digital intelligence

Multimedia

Image/video description

[Rui et al., 1999; Huang 2004]

Physical intelligence

Embodied AI, autonomous driving

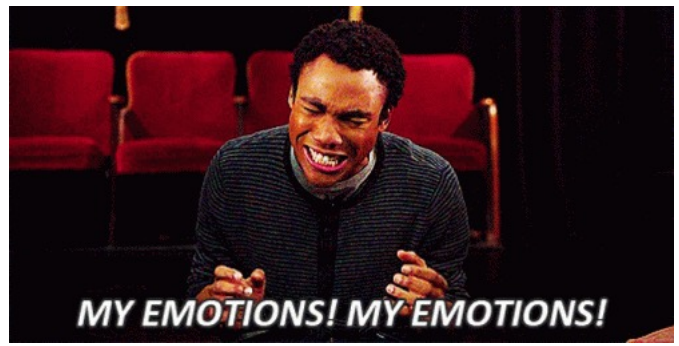
[Xu et al., 2017; Szot et al., 2021]

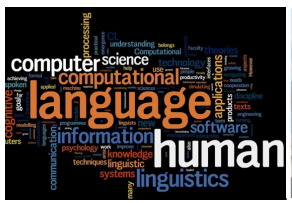
Social intelligence

Affective computing

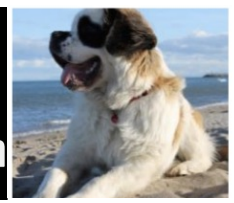
Human-AI interaction

[Picard 1997; Jaimes & Sebe 2007]

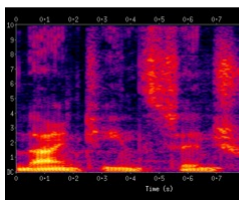




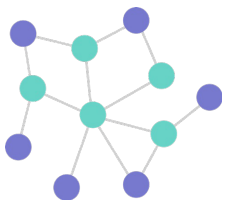
Language (written)



Image



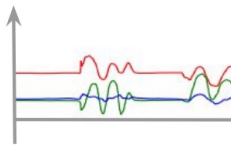
Audio



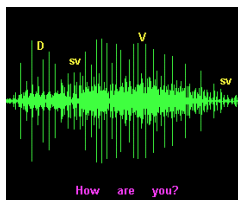
Graphs

Video

LIDAR

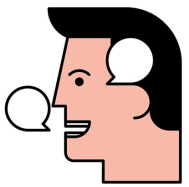


Sensors



Speech

Video (faces)



Language (spoken)

Multimodal Behaviors and Signals

Language

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Acoustic

- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual

- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Touch

- **Haptics**
- **Motion**

Physiological

- **Skin conductance**
- **Electrocardiogram**

Mobile

- **GPS location**
- **Accelerometer**
- **Light sensors**

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Behavioral Study of Multimodal

Language and gestures

David McNeill

"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

McGurk effect

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Behavioral Study of Multimodal

Language and gestures

David McNeill

"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

McGurk effect

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

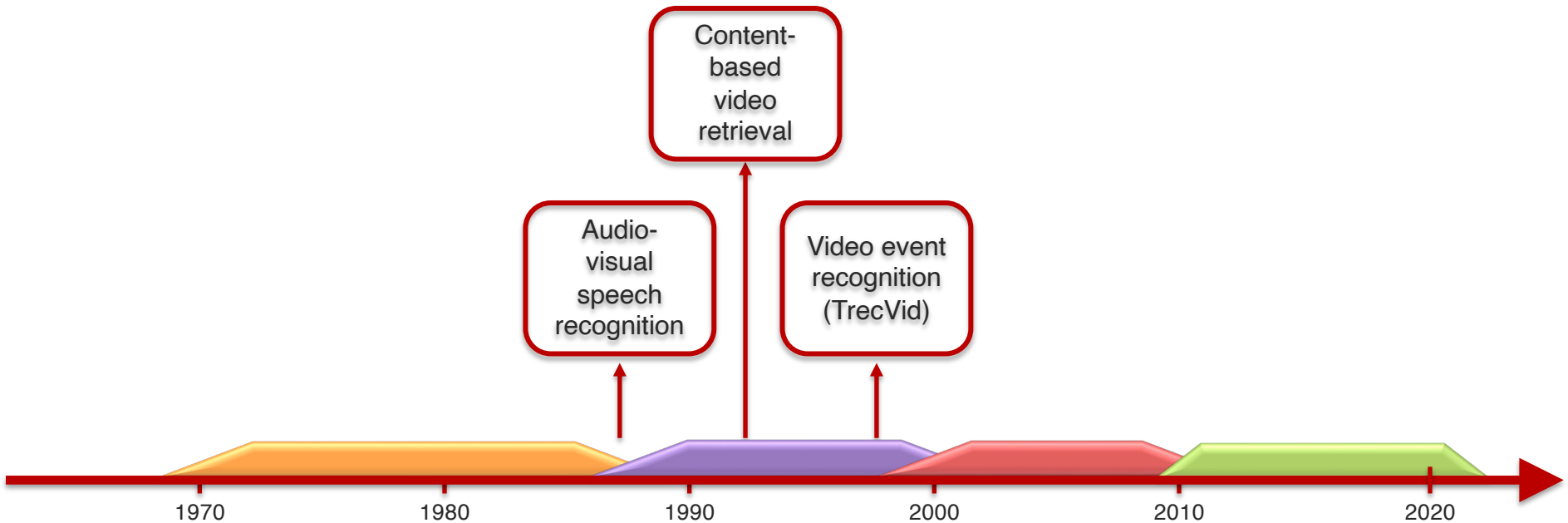
Prior Research in Multimodal

Four eras of multimodal research

- The "behavioral" era (1970s until late 1980s)
- The "computational" era (late 1980s until 2000)
- The "interaction" era (2000 - 2010)
- The "deep learning" era (2010s until …)

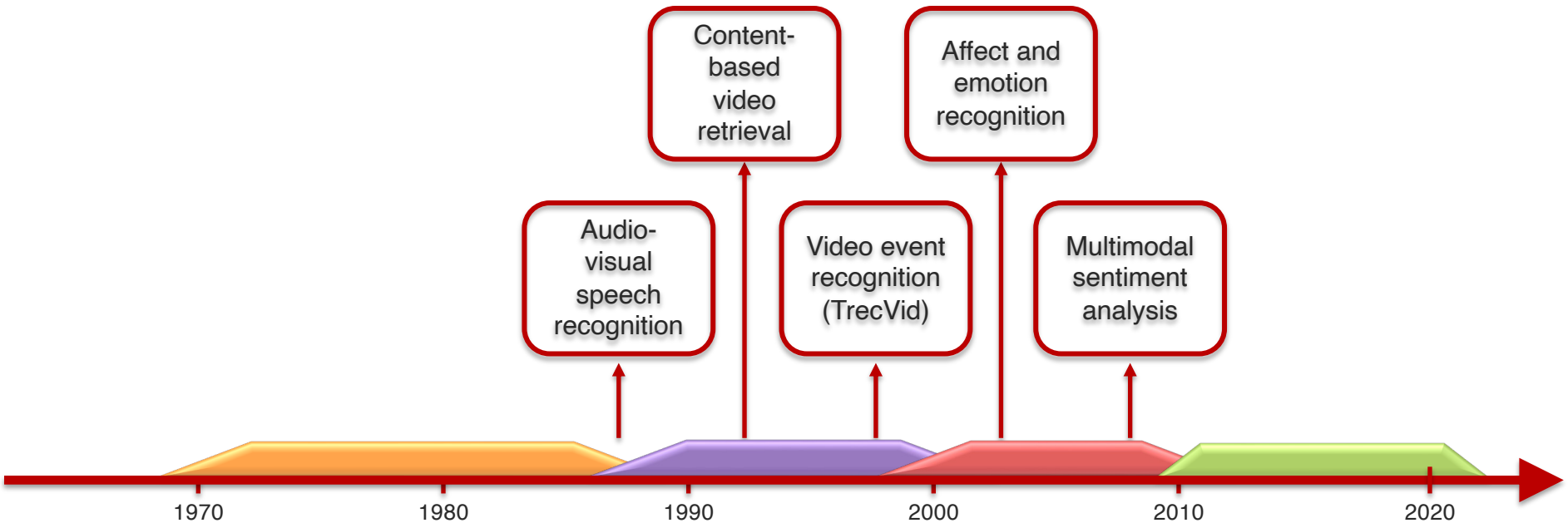
[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Research Tasks



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Research Tasks



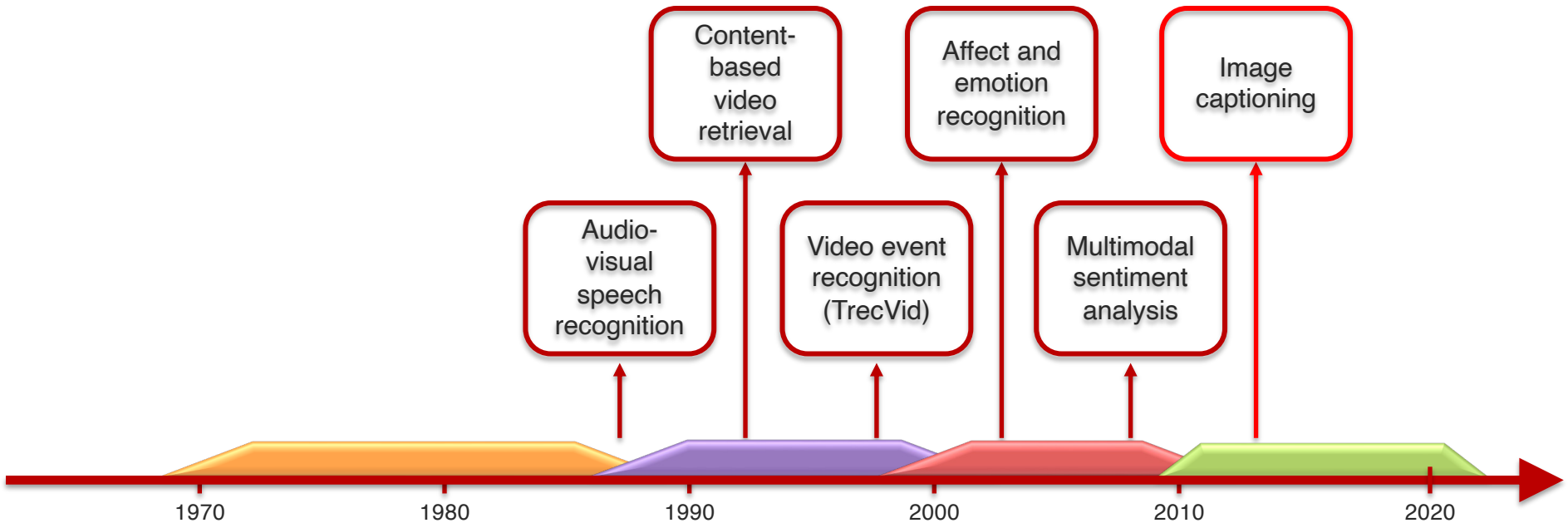
[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Research Tasks

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

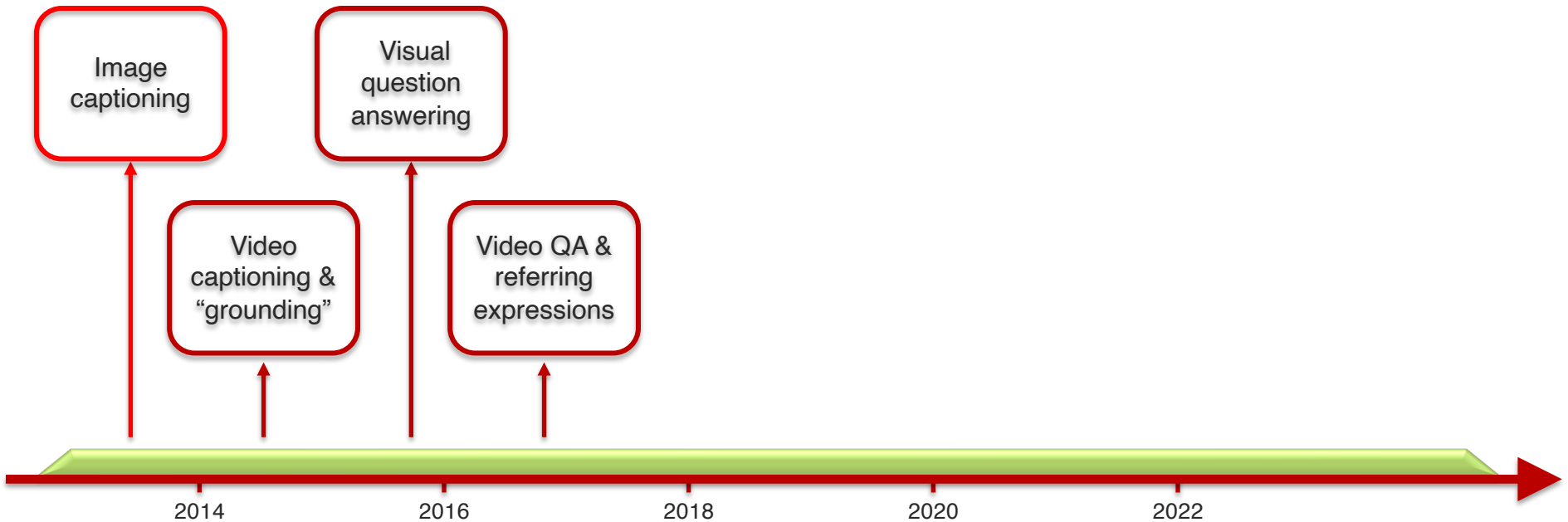
Multimodal Research Tasks

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

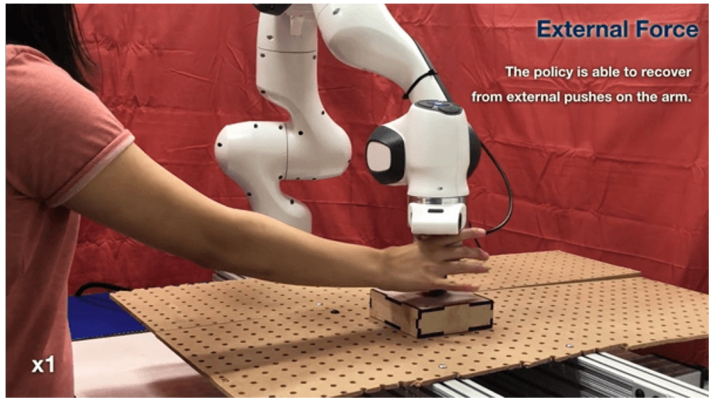
Is this person expecting company?
What is just under the tree?

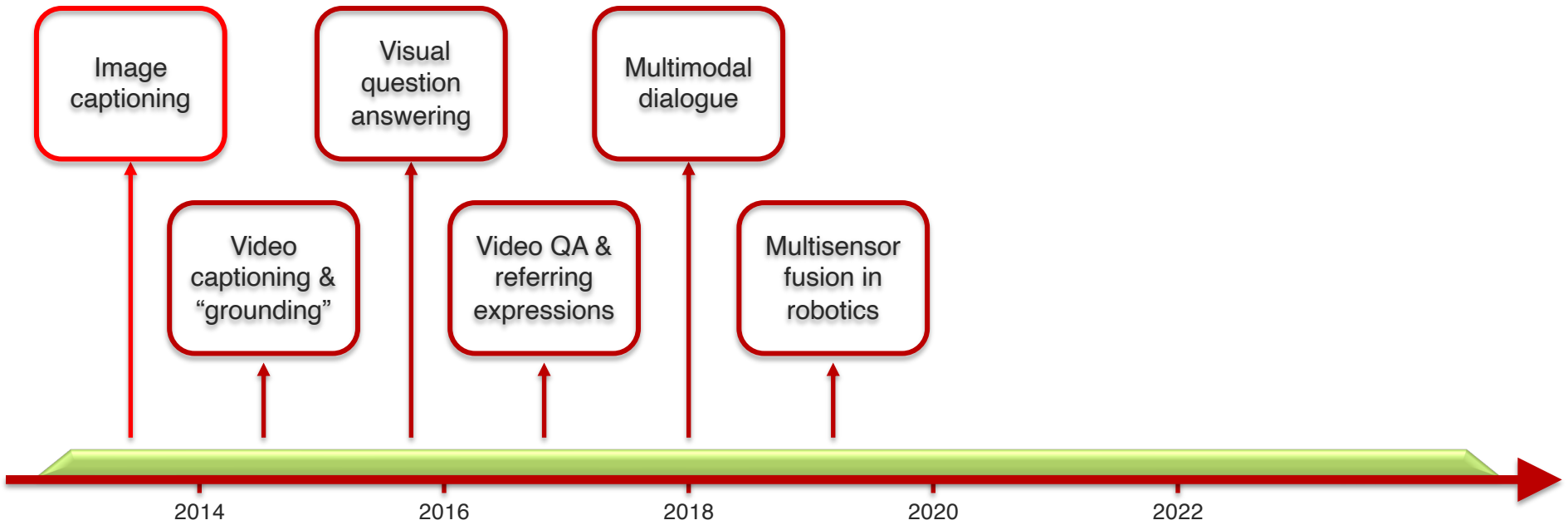
Does it appear to be rainy?
Does this person have 20/20 vision?



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

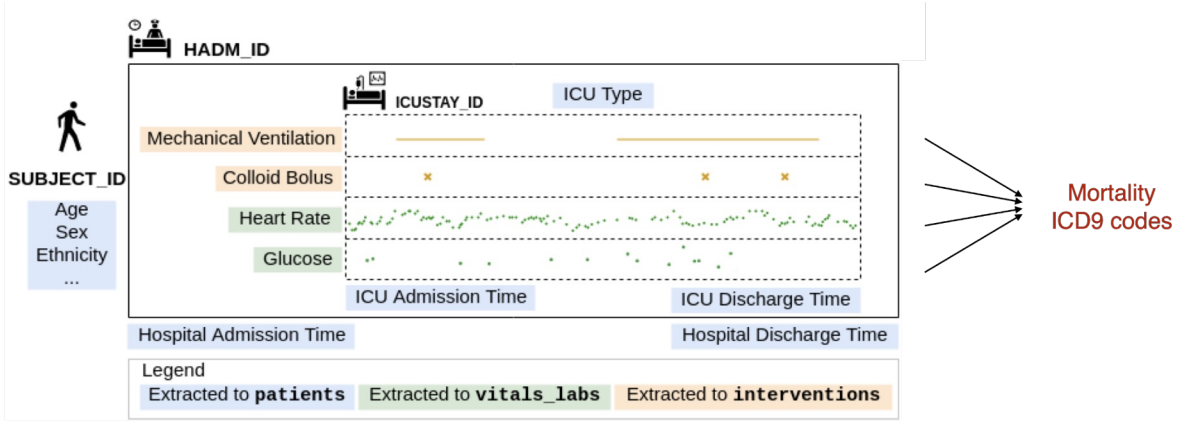
Multimodal Research Tasks

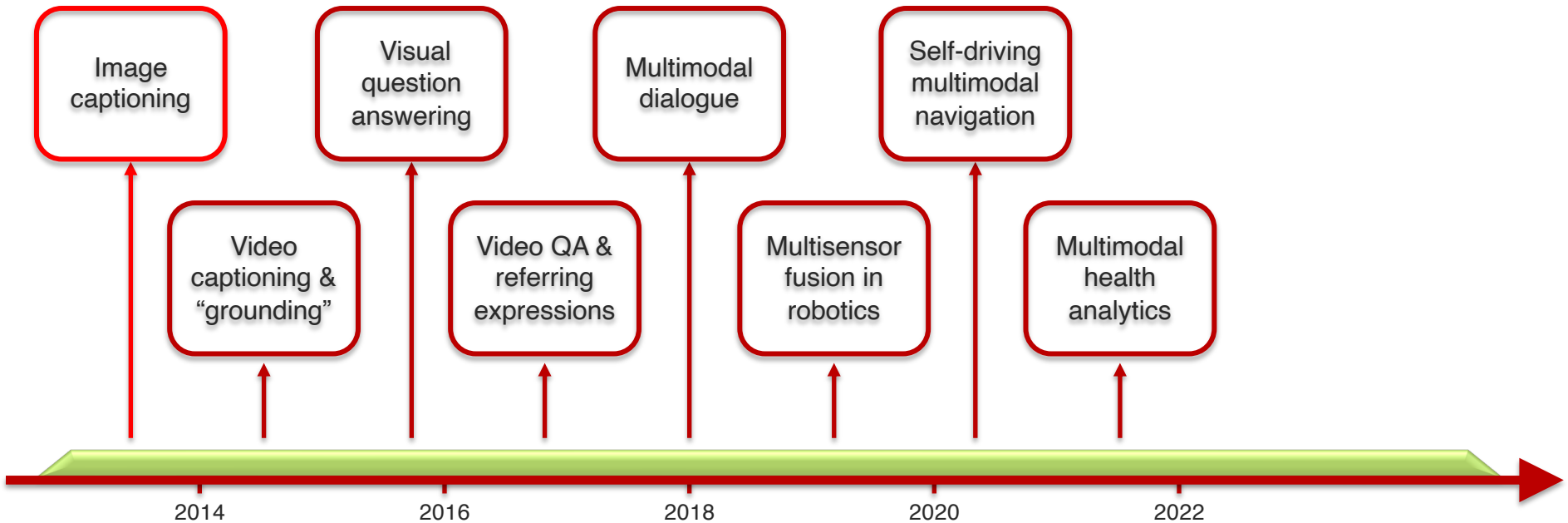




[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

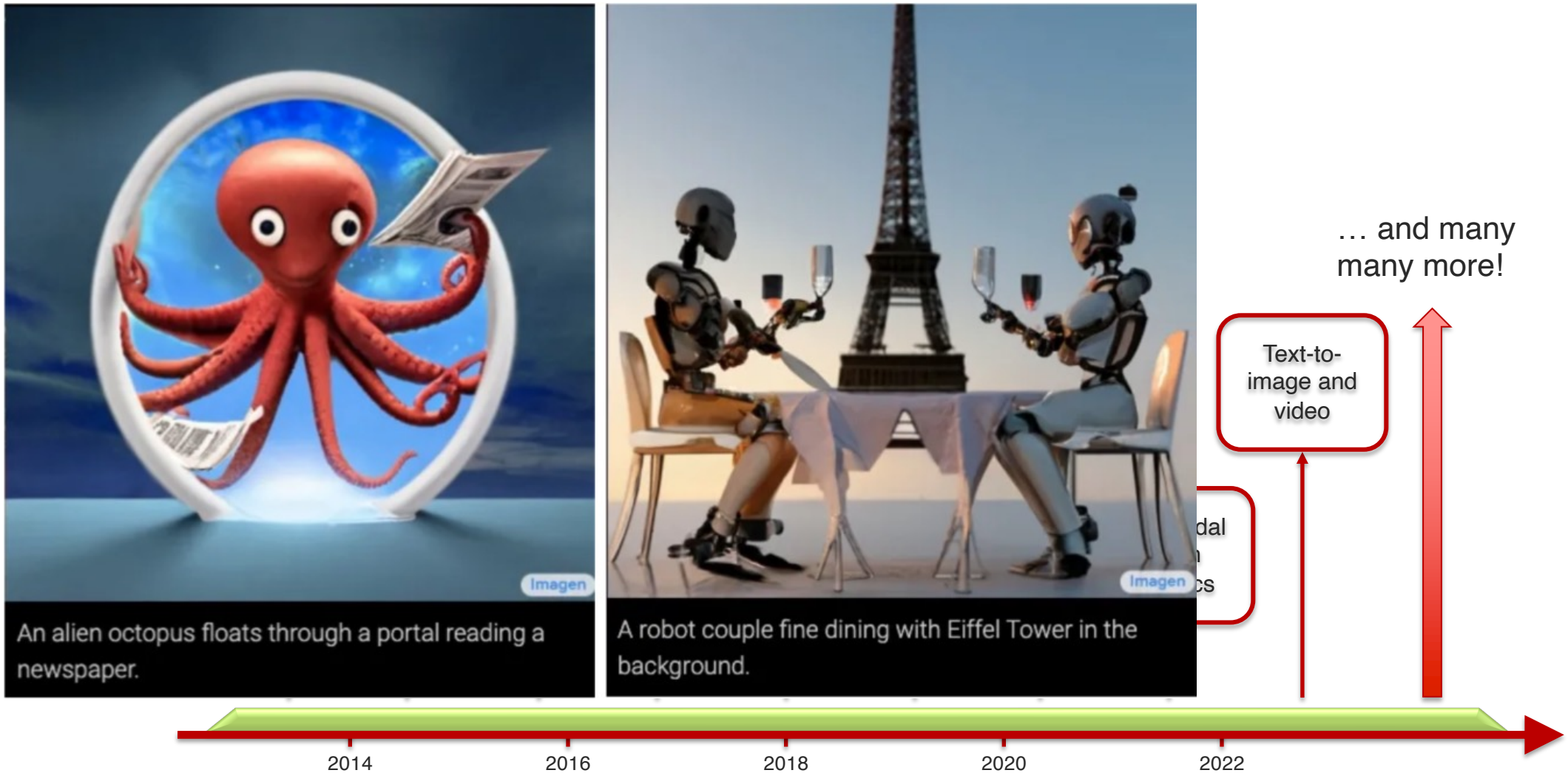
Multimodal Research Tasks





[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Research Tasks



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Challenges – Surveys, Tutorials and Courses

2016

Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

(Arxiv 2017, IEEE TPAMI journal, February 2019)

https://arxiv.org/abs/1705.09406

Tutorials: CVPR 2016, ACL 2016, ICMI 2016, …

Graduate-level courses:

Multimodal Machine Learning (11th edition)
https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

Advanced Topics in Multimodal Machine Learning
https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

2022

Foundations and Recent Trends in Multimodal ML

Paul Liang, Amir Zadeh, and Louis-Philippe Morency

- 6 core challenges
- 50+ taxonomic classes
- 600+ referenced papers

Tutorials: CVPR 2022, NAACL 2022, …

Updated graduate-level course:

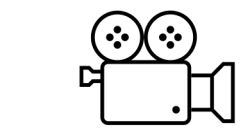
Multimodal Machine Learning (12th edition)
Fall 2022 semester

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

What is a Modality?

Modality

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, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

A research-oriented definition…

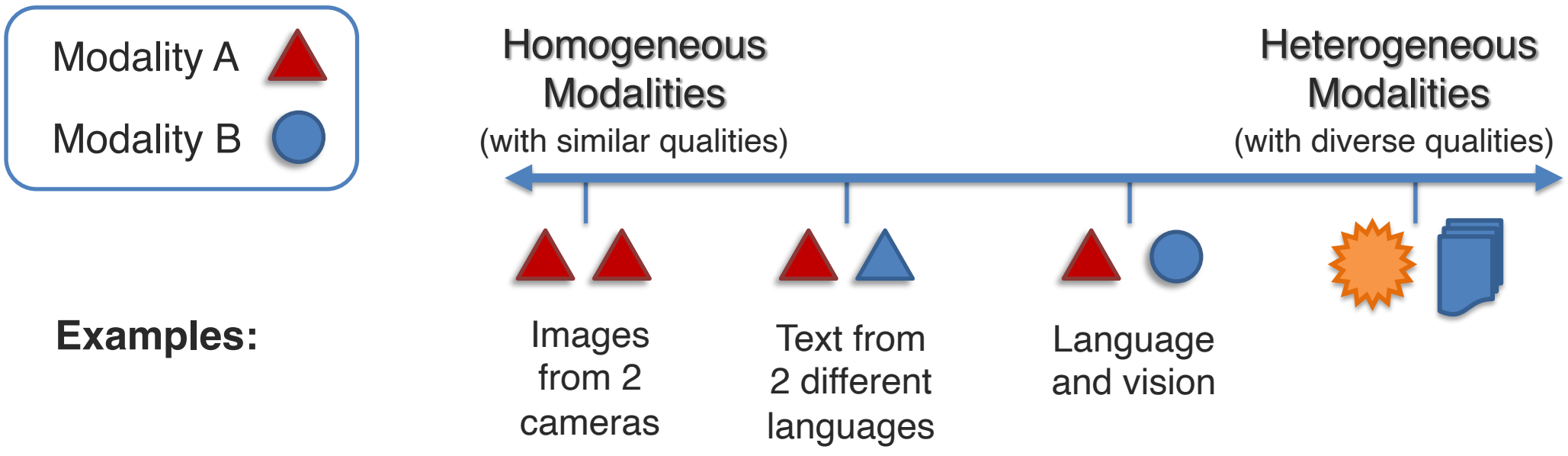
***Multimodal* is the science of**

heterogeneous and interconnected data

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Heterogeneous Modalities

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



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Dimensions of Heterogeneity

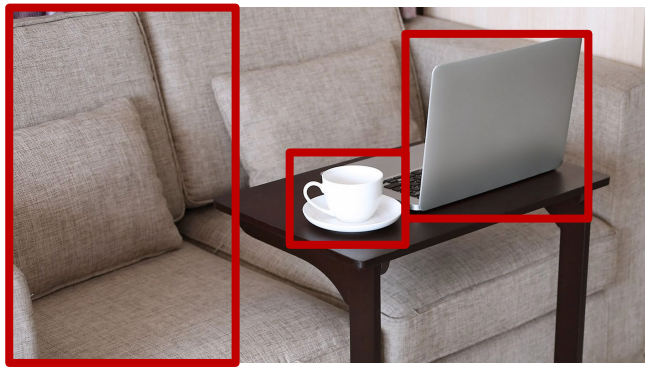
Information present in different modalities will often show diverse qualities, structures, and representations.

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

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. 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**.*

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

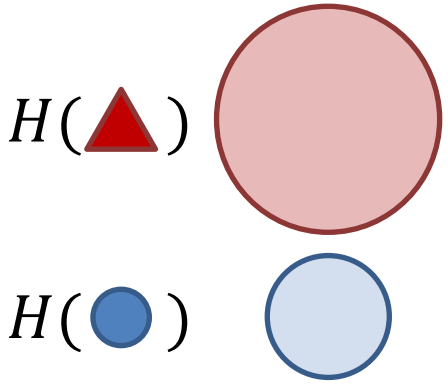
words per minute

Dimensions of Heterogeneity

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

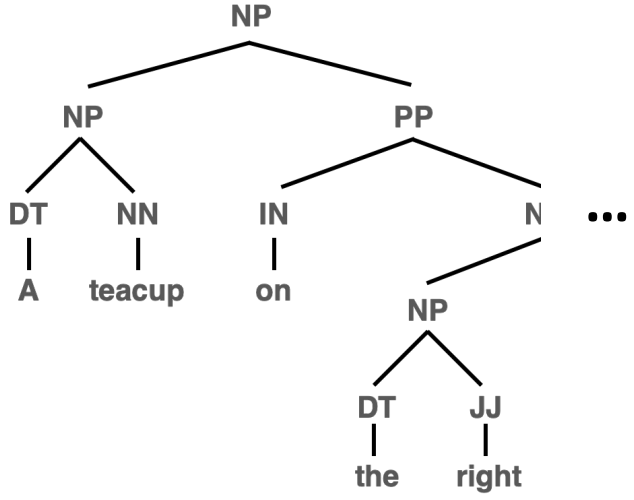
A teacup on the right of a laptop in a clean room.

3 **Information:** entropy and density

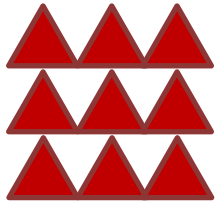


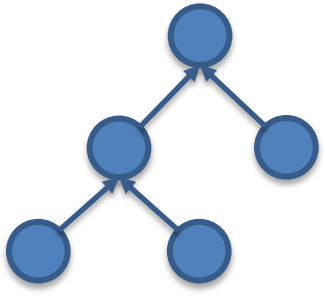
Dimensions of Heterogeneity

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.

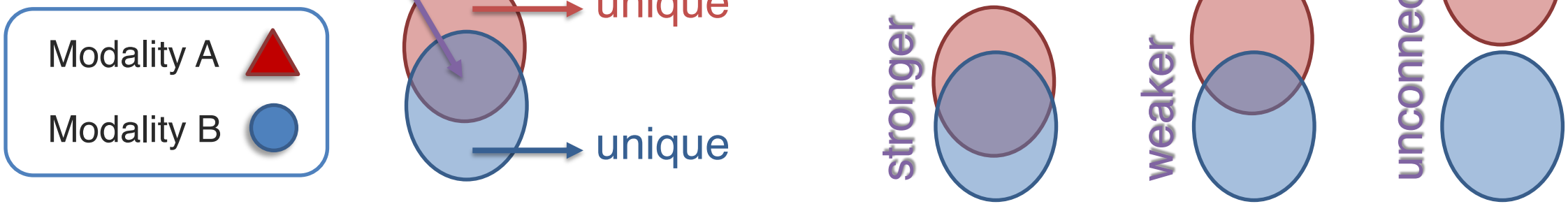
5 **Noise:** uncertainty, signal-to-noise ratio, missing data

teacup → teacip

right → rihjt

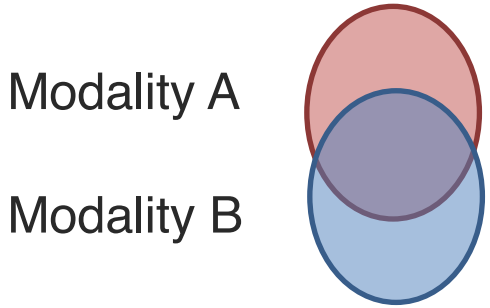
Interconnected Modalities

Connected: Shared information that relates modalities



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



① **Connections**

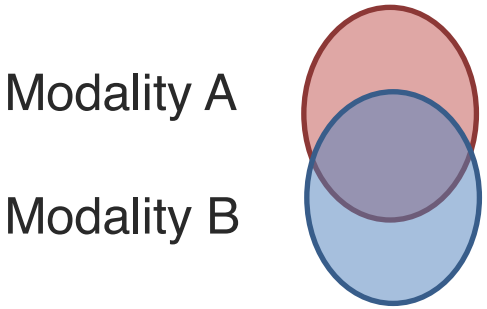
Shared information that relates modalities.

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

⟷ ***teacup***

⟷ ***laptop***

Interconnected Modalities

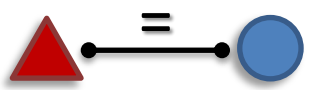


① **Connections**

Shared information that relates modalities.

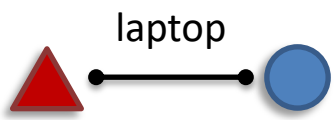
Statistical → **Semantic**

Association



e.g., correlation, co-occurrence

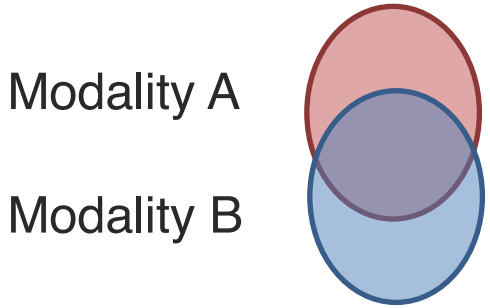
Correspondence



e.g., grounding

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



① **Connections**

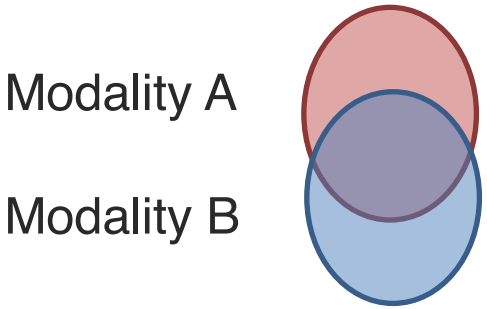
Shared information that relates modalities.

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

clean

room

Interconnected Modalities

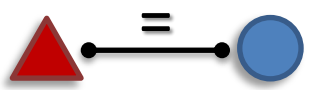


① **Connections**

Shared information that relates modalities.

Statistical

Association



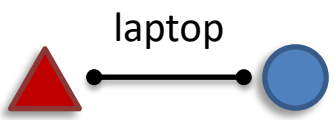
e.g., correlation, co-occurrence

Dependency

e.g., causal, temporal

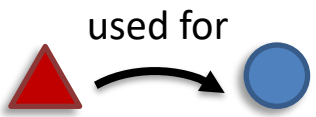
Semantic

Correspondence



e.g., grounding

Relationship

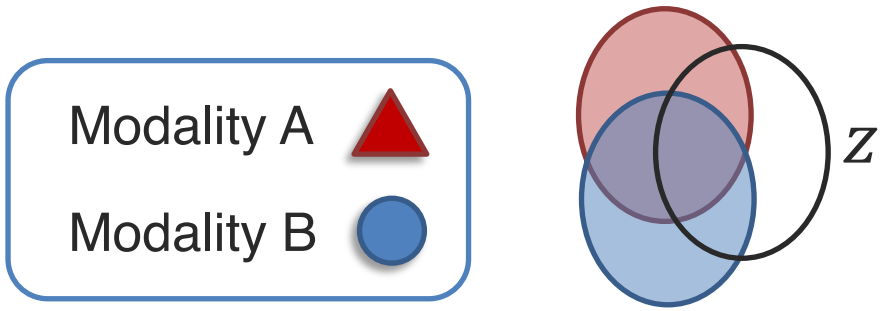


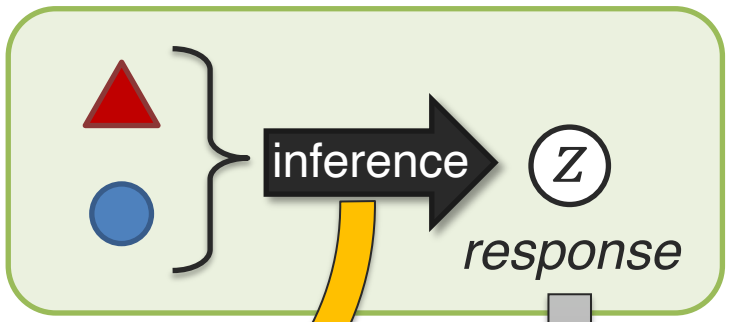
e.g., function

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities

Interacting: process affecting each modality, creating new response

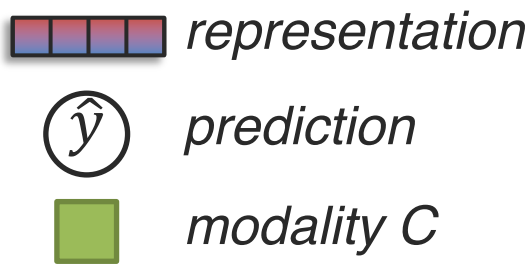




Interactions happen during inference!

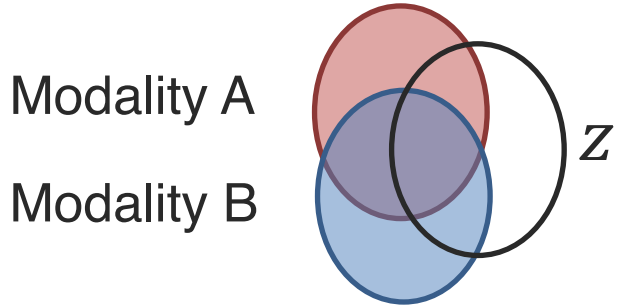
"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



② **Cross-modal interactions**

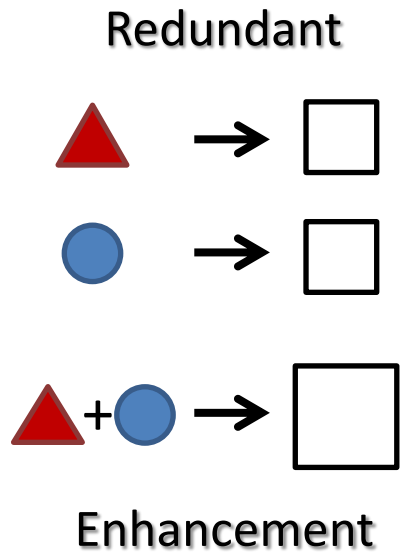
Modality elements often interact during inference.

Is this indoors?

inference → *Yes!*

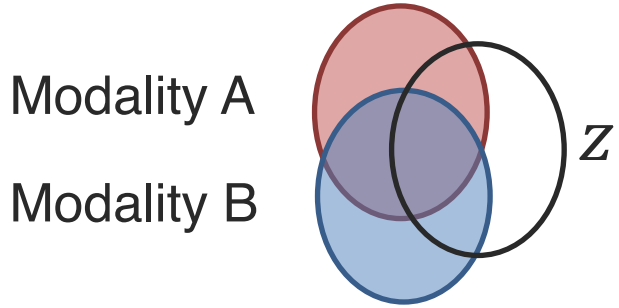
A teacup on the right of a laptop in a clean room.

inference → *Yes!*



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



② **Cross-modal interactions**

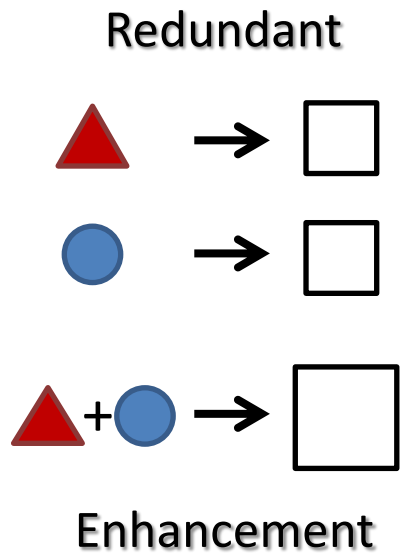
Modality elements often interact during inference.

Is this indoors?

inference → ***Yes!***

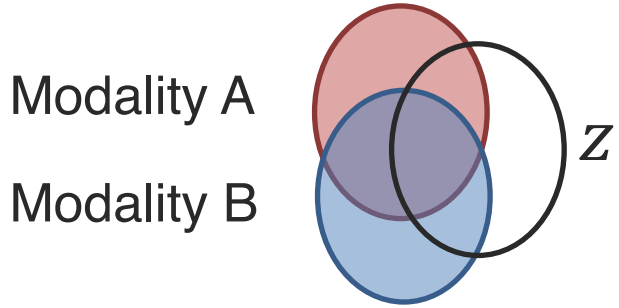
A teacup on the right of a laptop in a clean room.

inference → ***Yes!***



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



② **Cross-modal interactions**

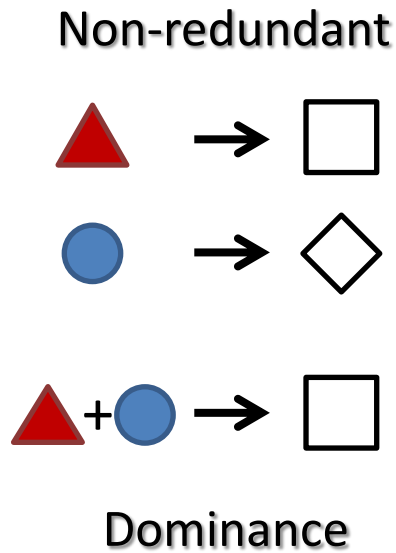
Modality elements often interact during inference.

Is this a living room?

inference → *Yes!*

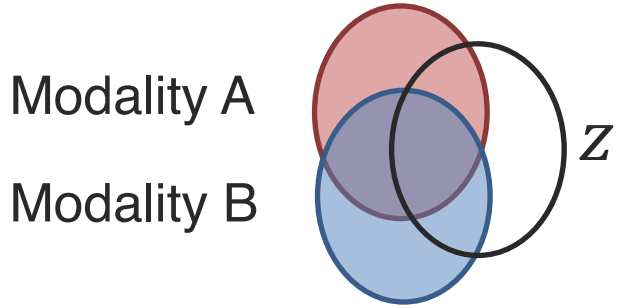
inference → *No, probably study room.*

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



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



② **Cross-modal interactions**

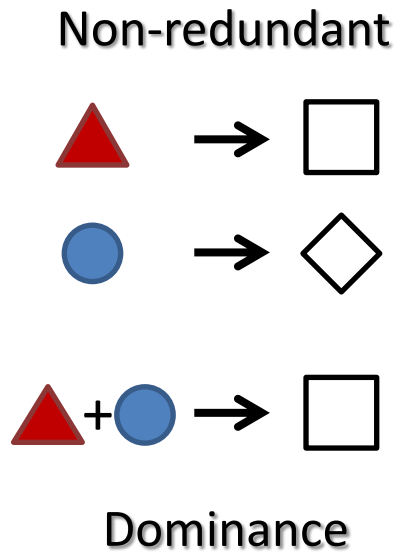
Modality elements often interact during inference.

Is this a living room?

inference → ***Yes!***

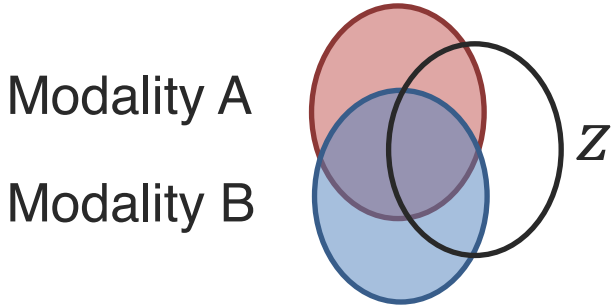
A teacup on the right of a laptop in a clean room.

inference → *No, probably study room.*



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Interconnected Modalities



② Cross-modal interactions

Modality elements often interact during inference.

Should I work here?

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

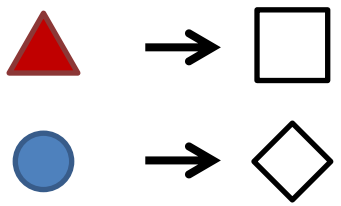
inference

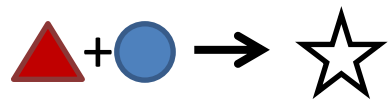
Maybe? Comfy sofa but table's too small.

inference

Maybe? Clean and there's tea.

Non-redundant

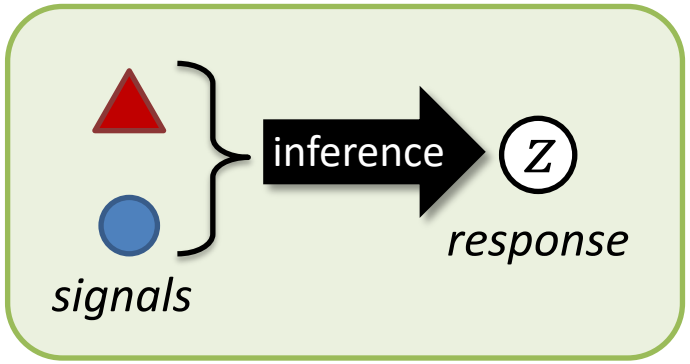




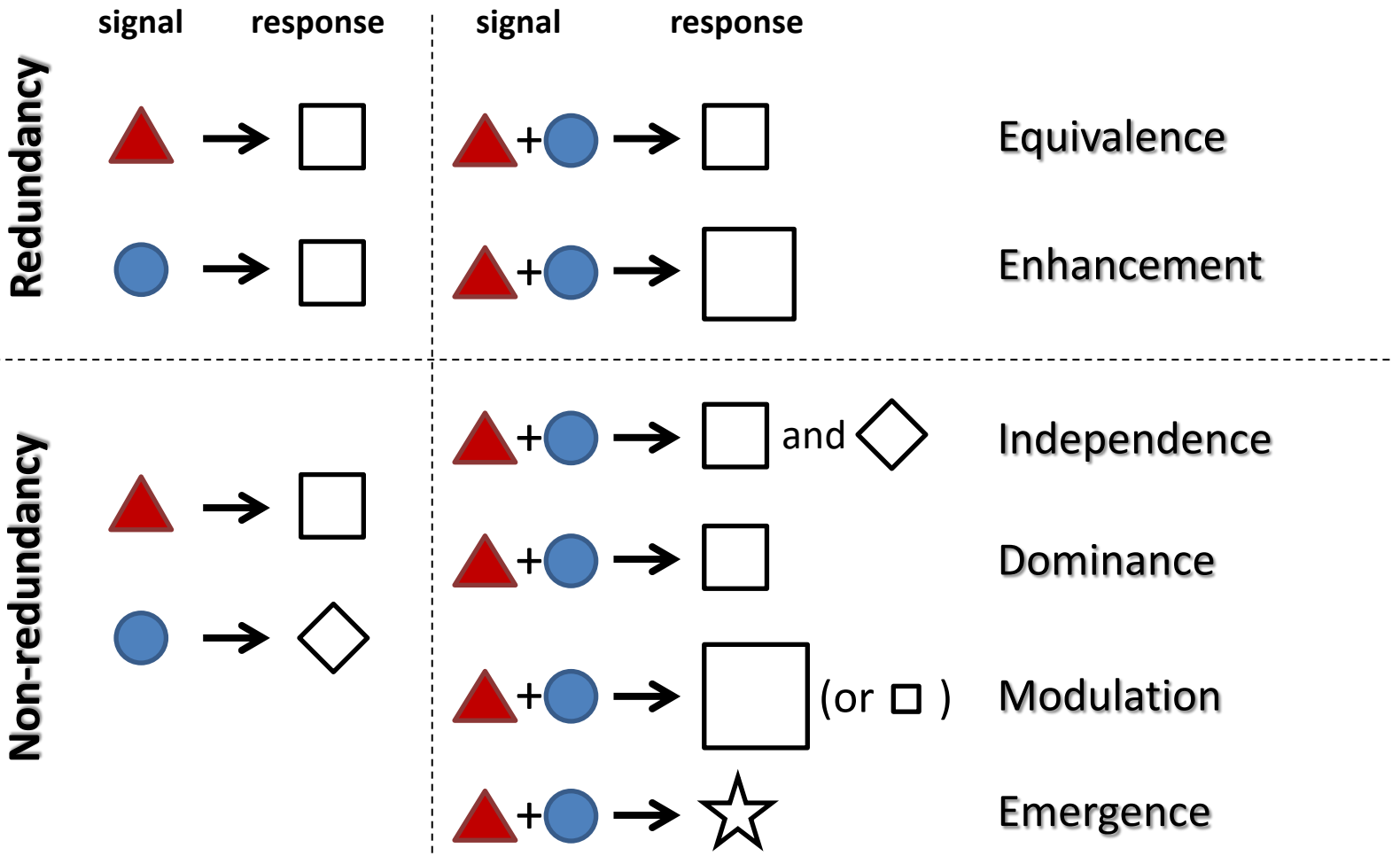
Emergence

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Cross-modal Interactions – A Behavioral Science View

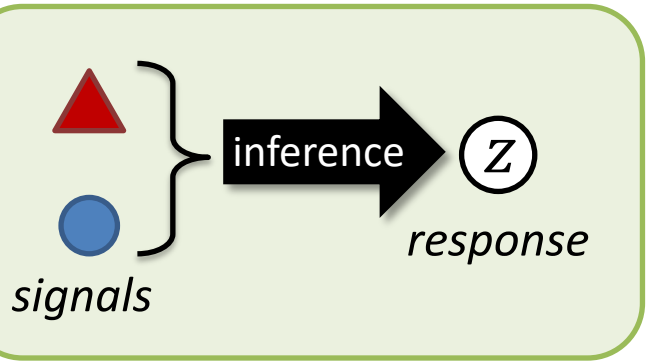


Multimodal Communication

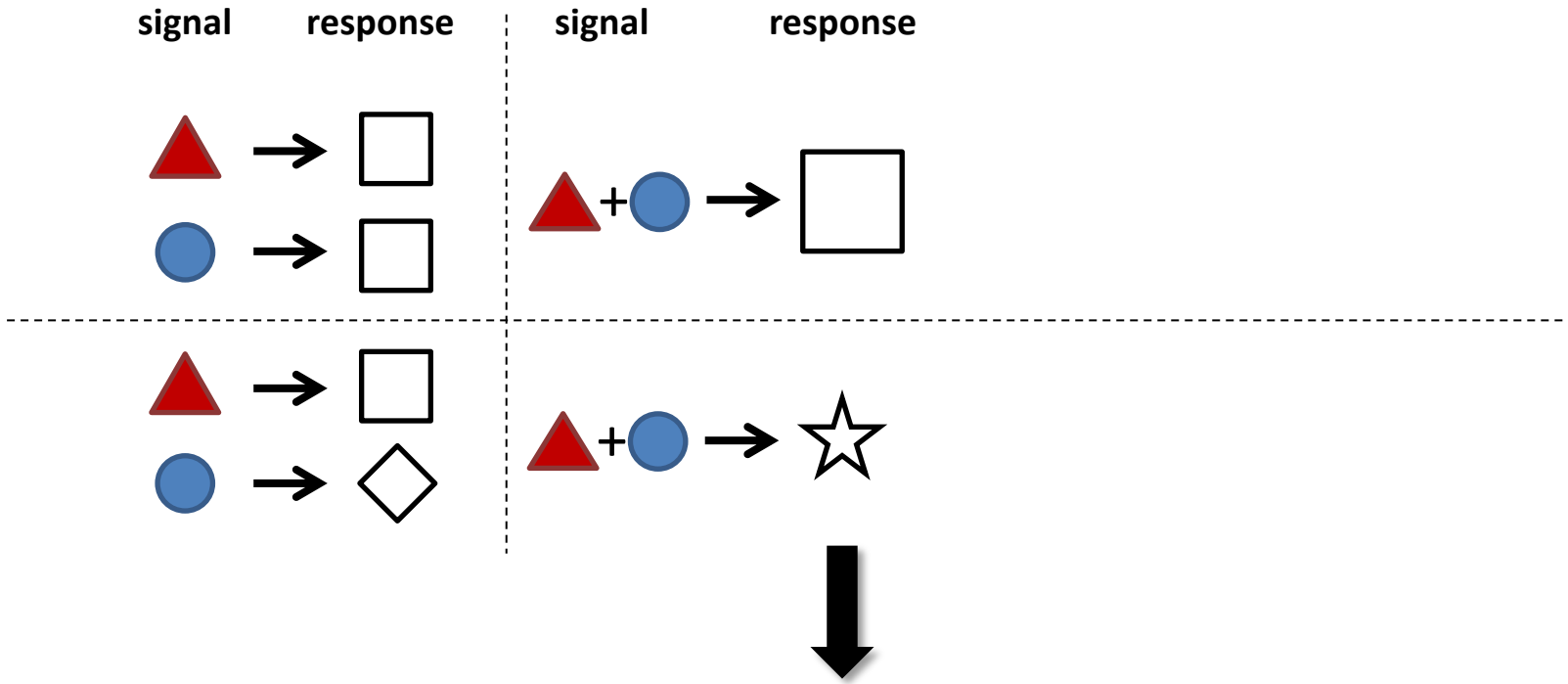


[Partan and Marler. Communication Goes Multimodal. Science 1999]

Dimensions of Cross-modal Interactions



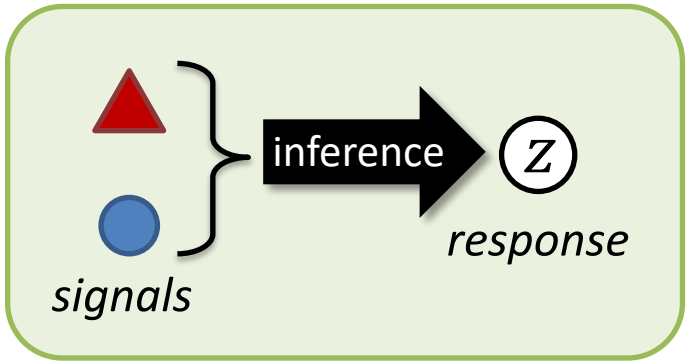
What are the dimensions for **digitally-represented** modalities?



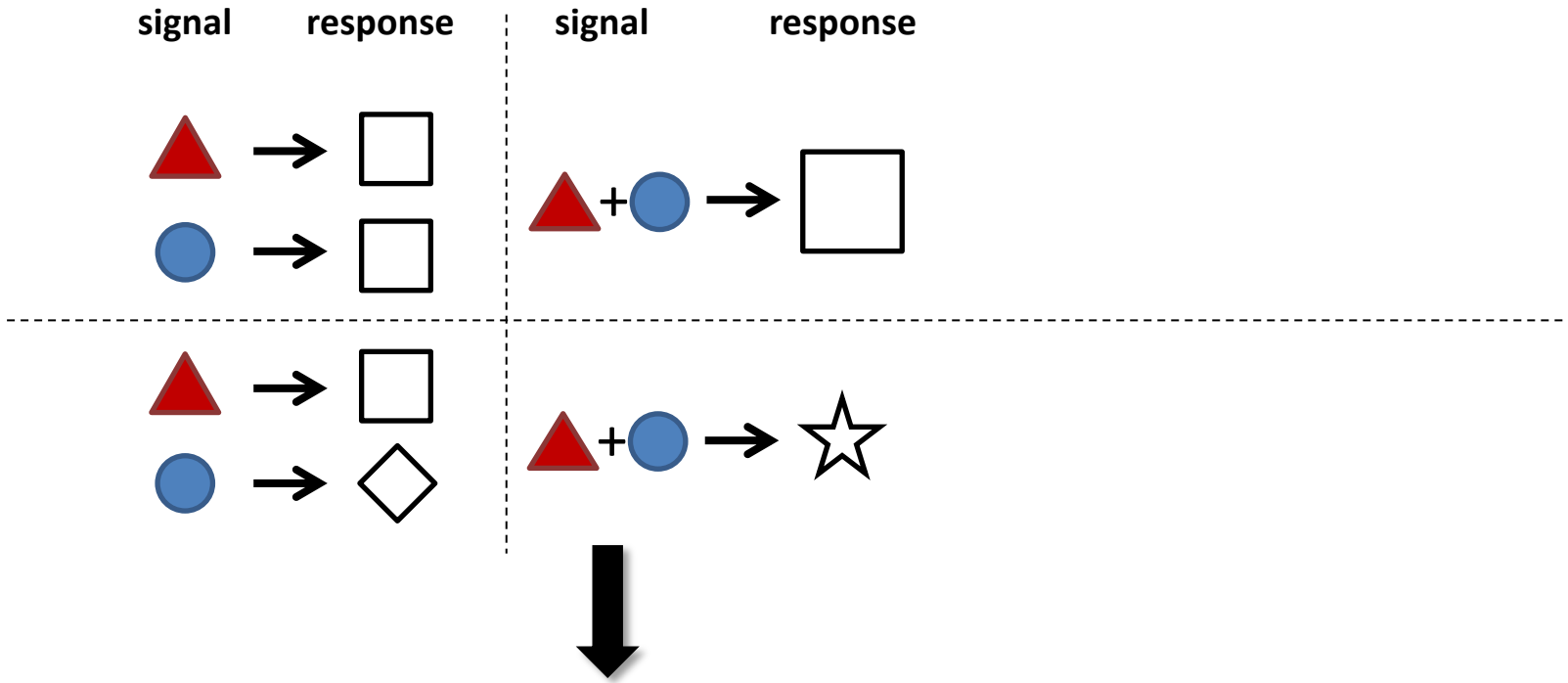
1 **Response:** Redundant, non-redundant, dominance, emergence…

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Dimensions of Cross-modal Interactions



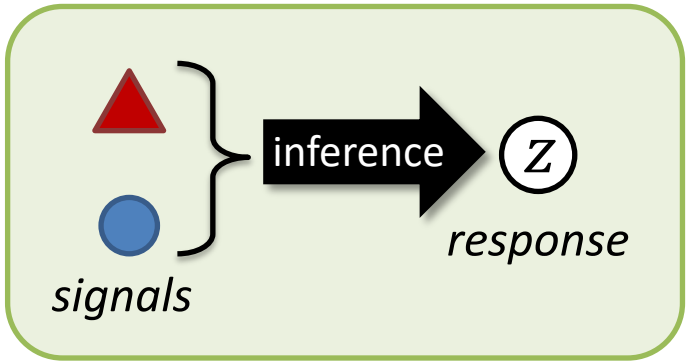
What are the dimensions for **digitally-represented** modalities?



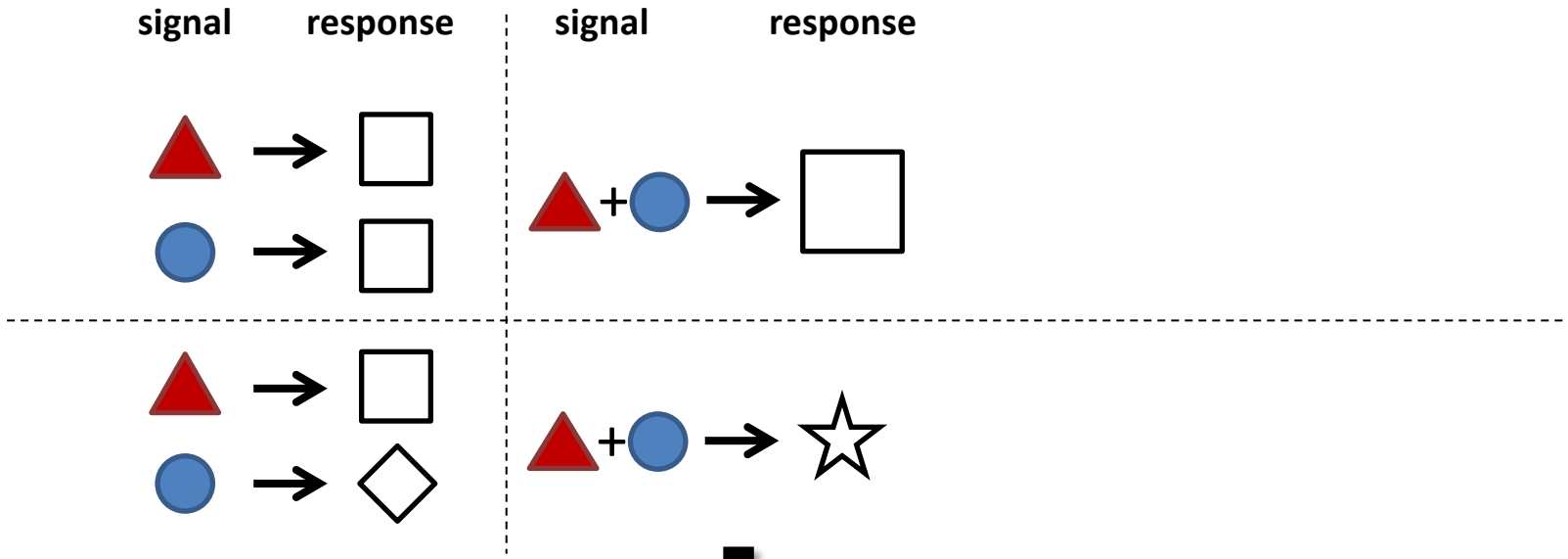
2 **Inputs:** Bimodal, trimodal, high-modal

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Dimensions of Cross-modal Interactions



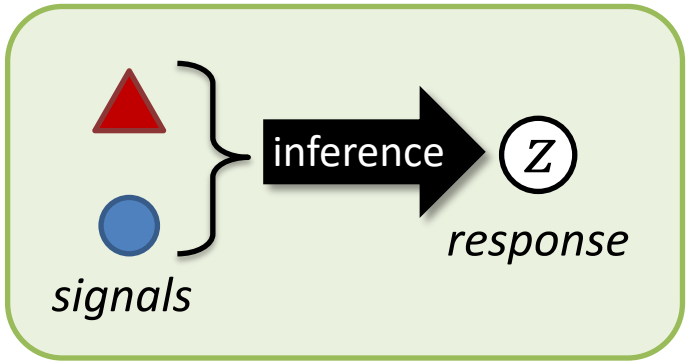
What are the dimensions for **digitally-represented** modalities?



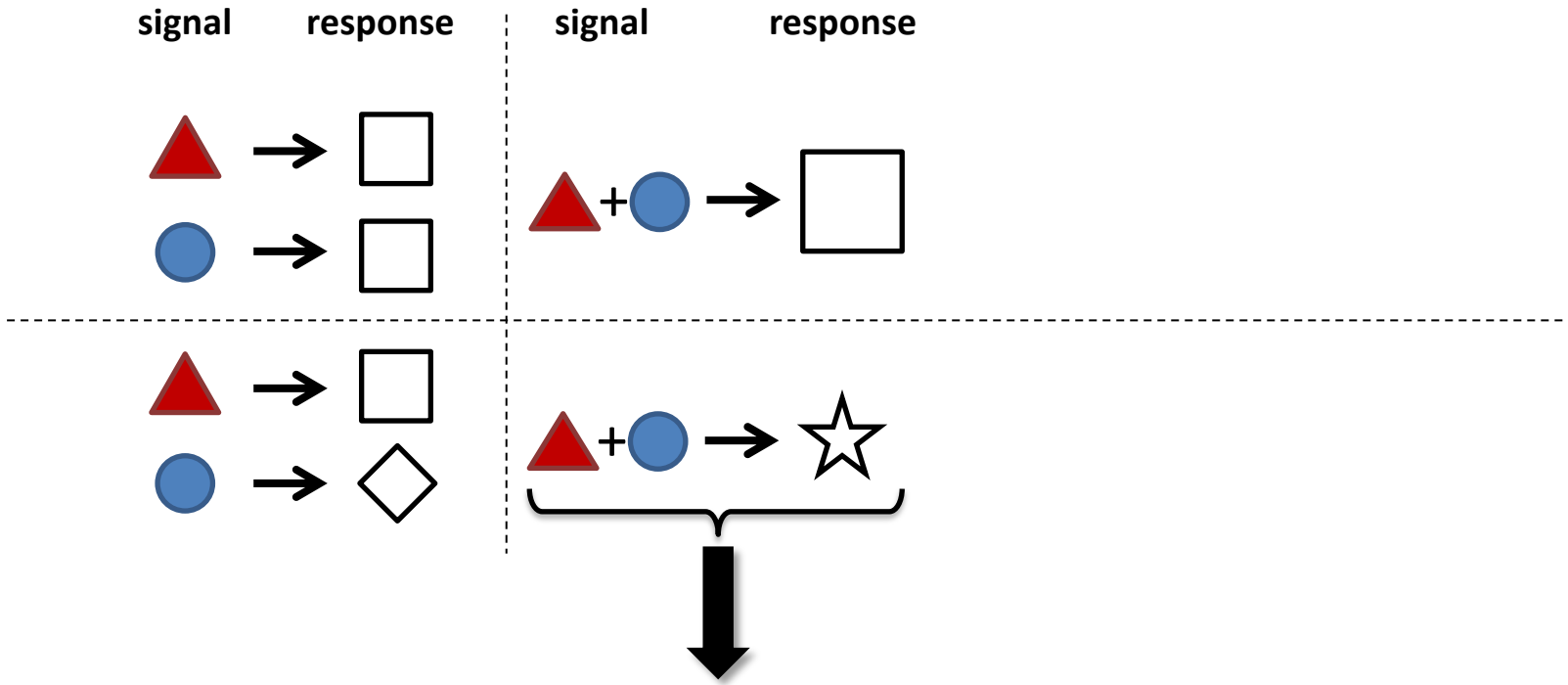
3 **Mechanics:** Additive, multiplicative, non-additive, causal, logical, …

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Dimensions of Cross-modal Interactions



What are the dimensions for **digitally-represented** modalities?



4 **Context:** Task relevance, context dependence

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

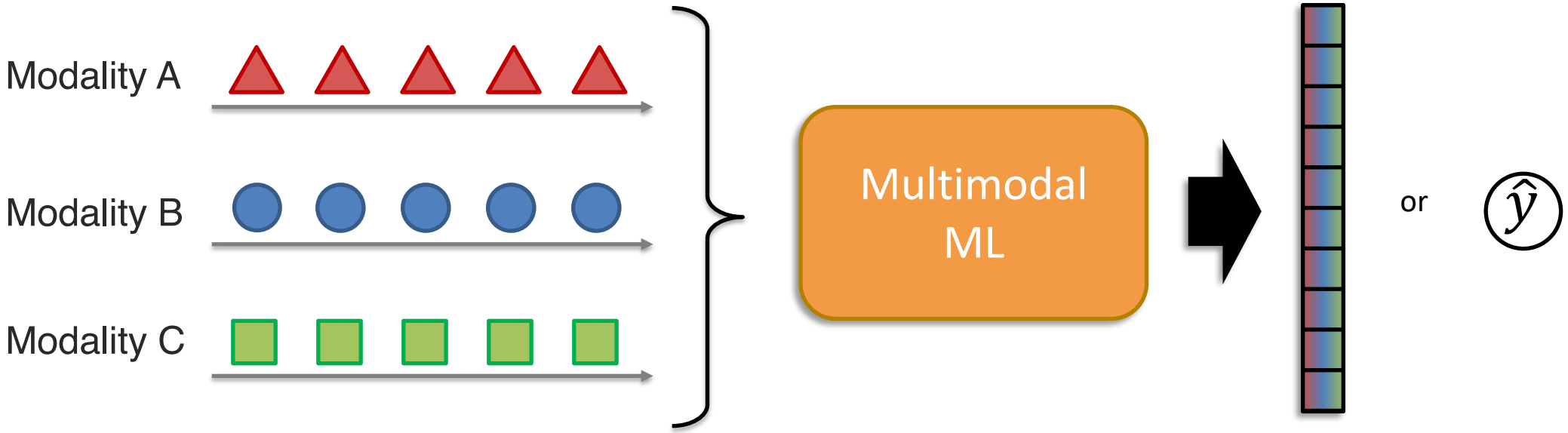
What is Multimodal?

***Multimodal* is the science of**

heterogeneous and interconnected data ☺

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Machine Learning



[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Multimodal Machine Learning

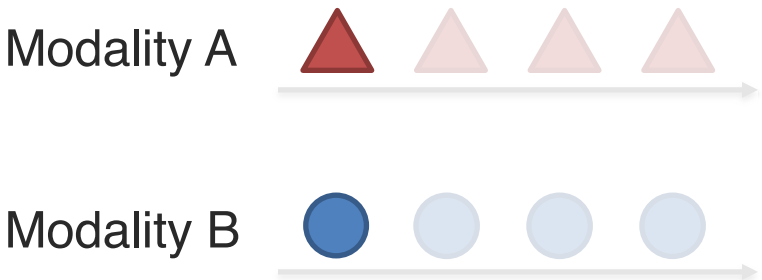
*What are the **core multimodal technical challenges,***

understudied in conventional machine learning?

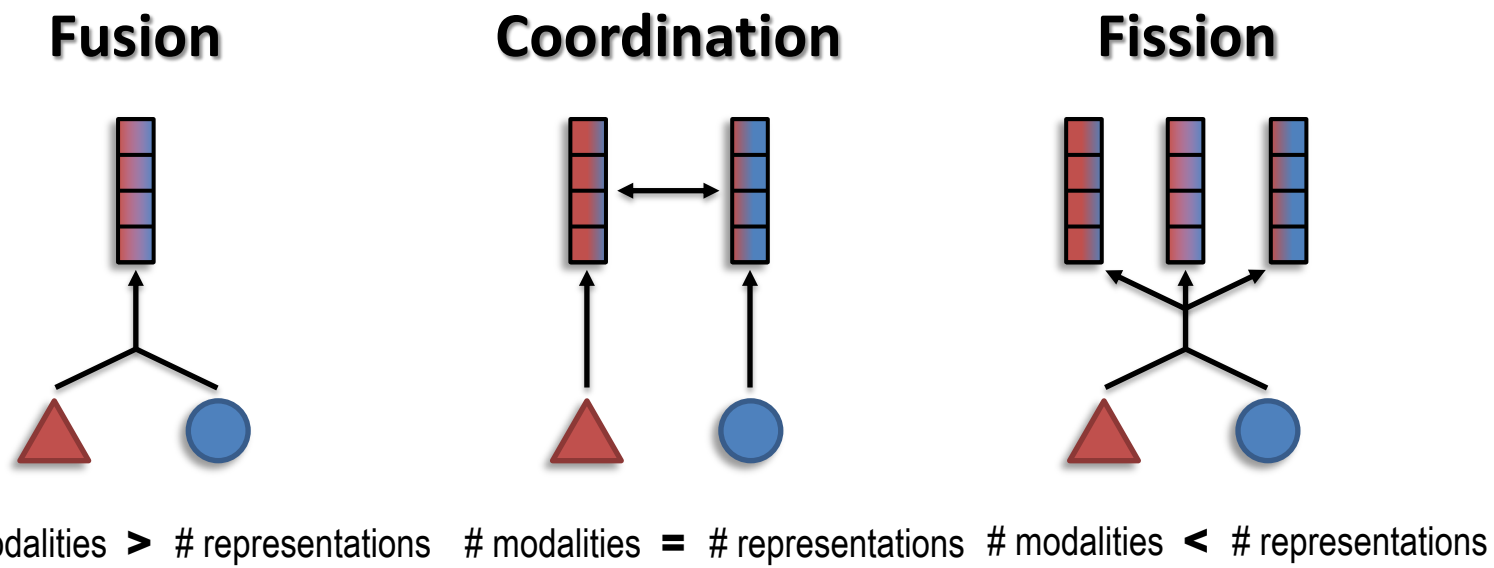
[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Challenge 1: Representation

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



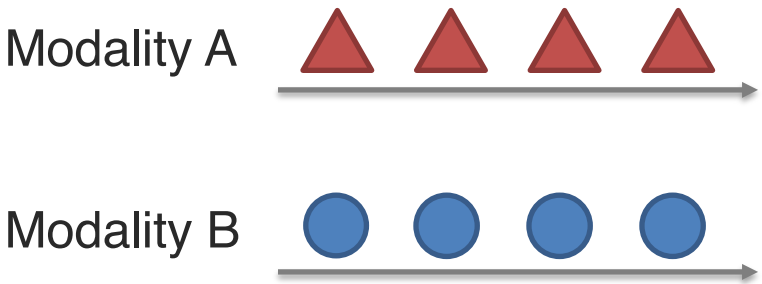
Sub-challenges:



[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

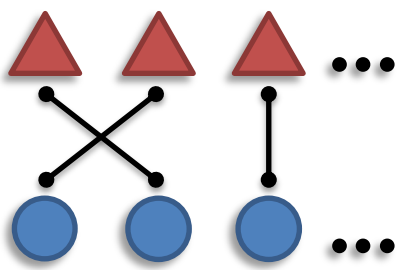
Challenge 2: Alignment

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



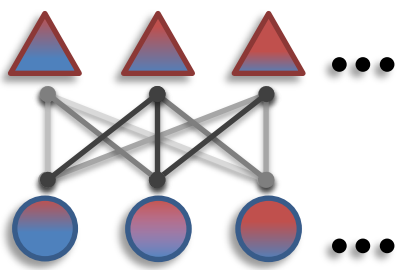
Sub-challenges:

Connections



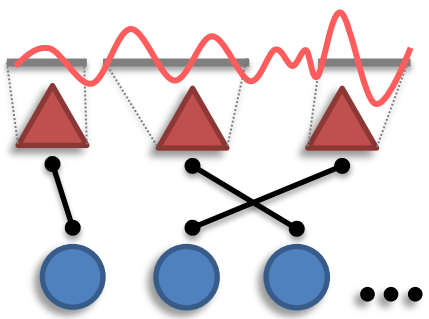
Explicit alignment
(e.g., grounding)

Contextualized Representation



Implicit alignment
+ representation

Segmentation

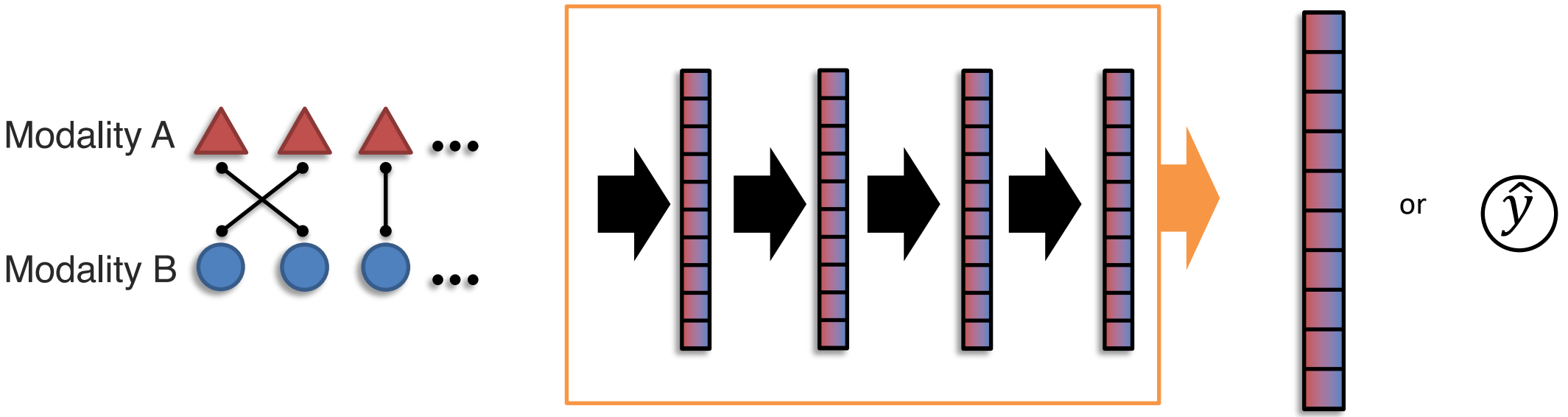


Granularity of individual elements

[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 3: Reasoning

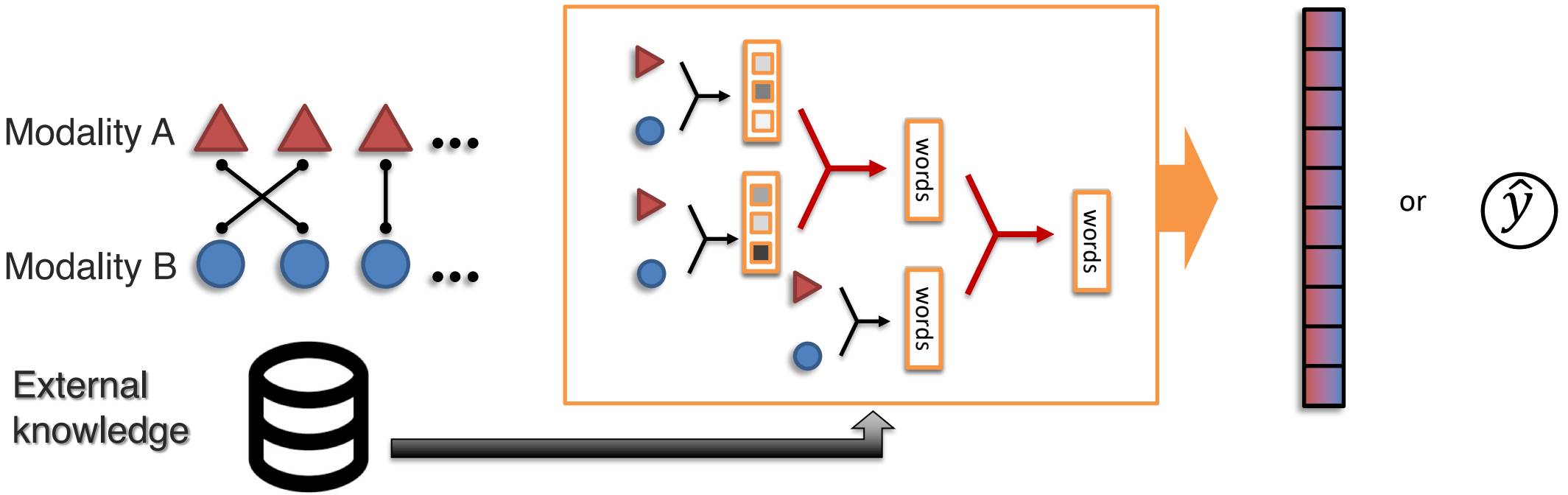
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 3: Reasoning

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



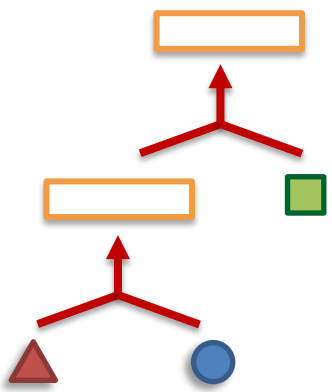
[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 3: Reasoning

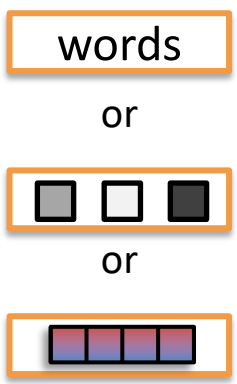
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

Sub-challenges:

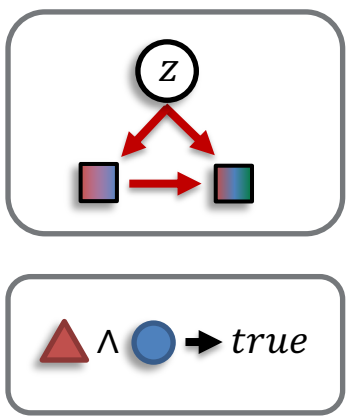
Structure Modeling



Intermediate concepts



Inference Paradigm



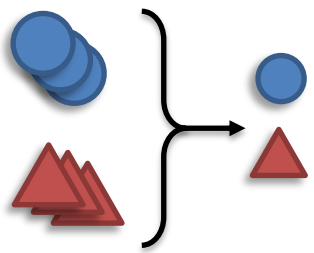
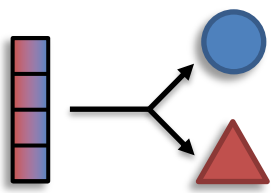
External Knowledge

[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 4: Generation

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

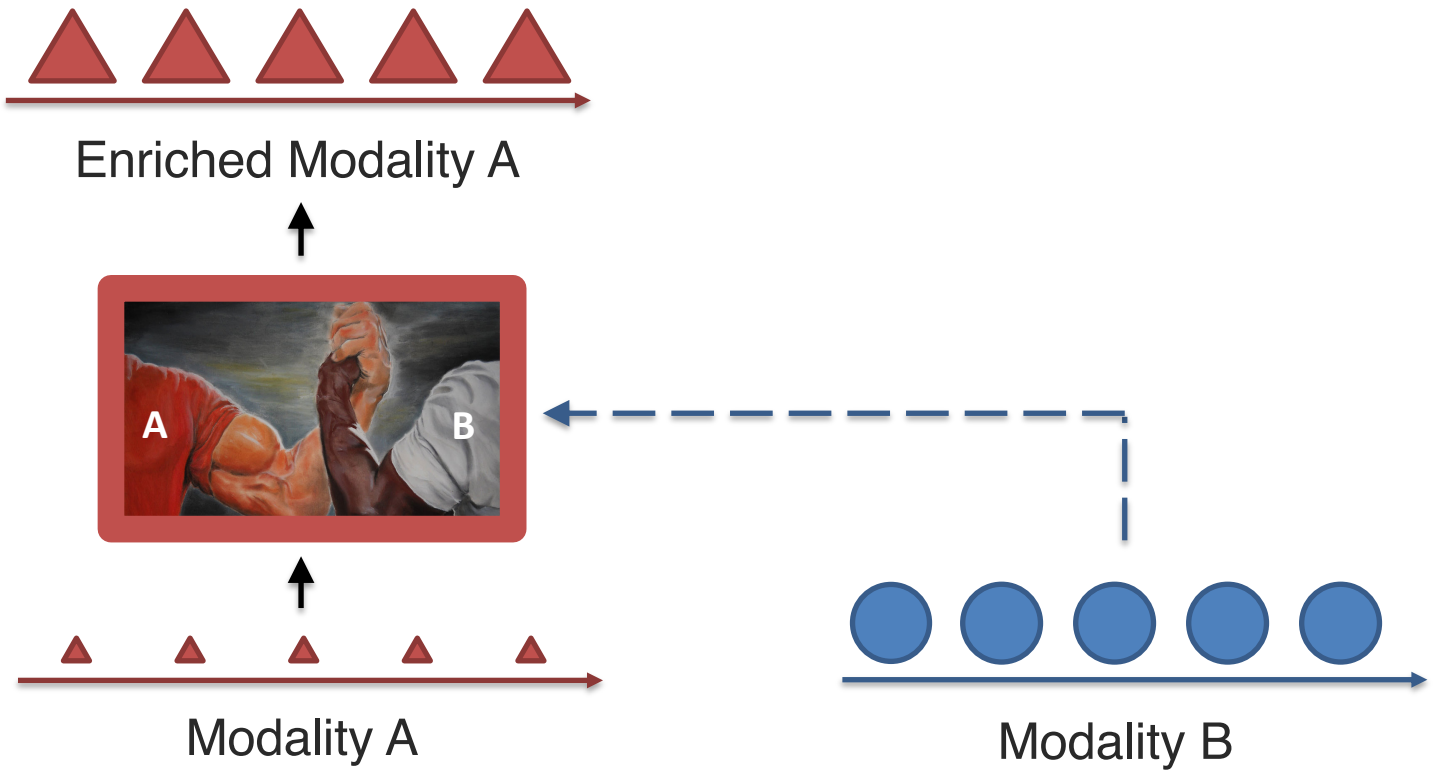
Sub-challenges:

	Summarization	Translation	Creation
			
Information: (content)	Reduction	Maintenance	Expansion
	□ > □	□ = □	□ < □

[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 5: Transference

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



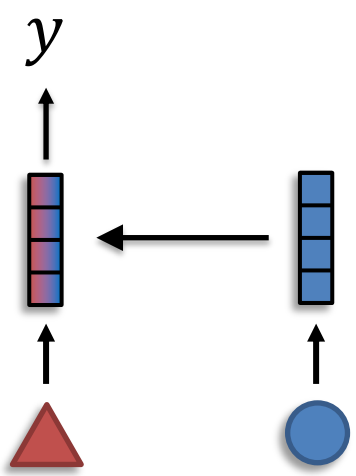
[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 5: Transference

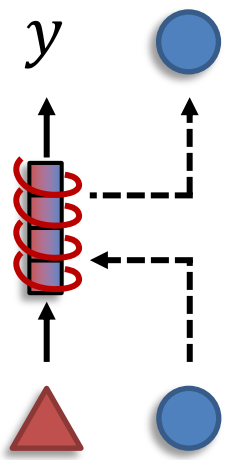
Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources.

Sub-challenges:

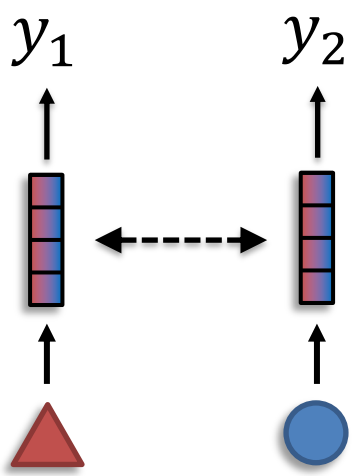
Transfer



Co-learning



Model Induction



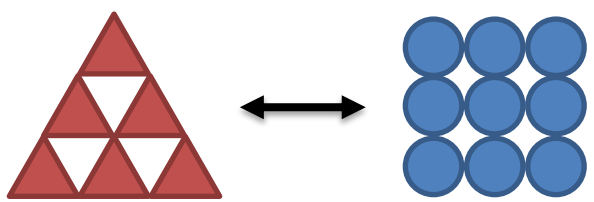
[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Challenge 6: Quantification

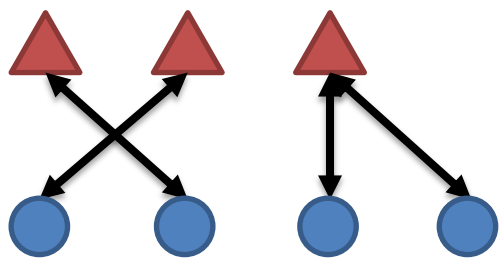
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions, and the multimodal learning process.

Sub-challenges:

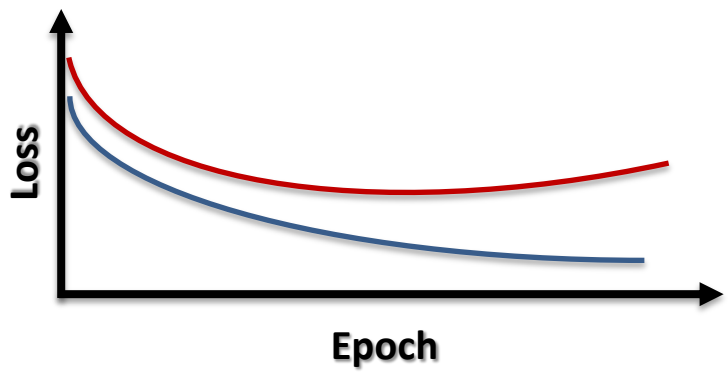
Heterogeneity



Interactions

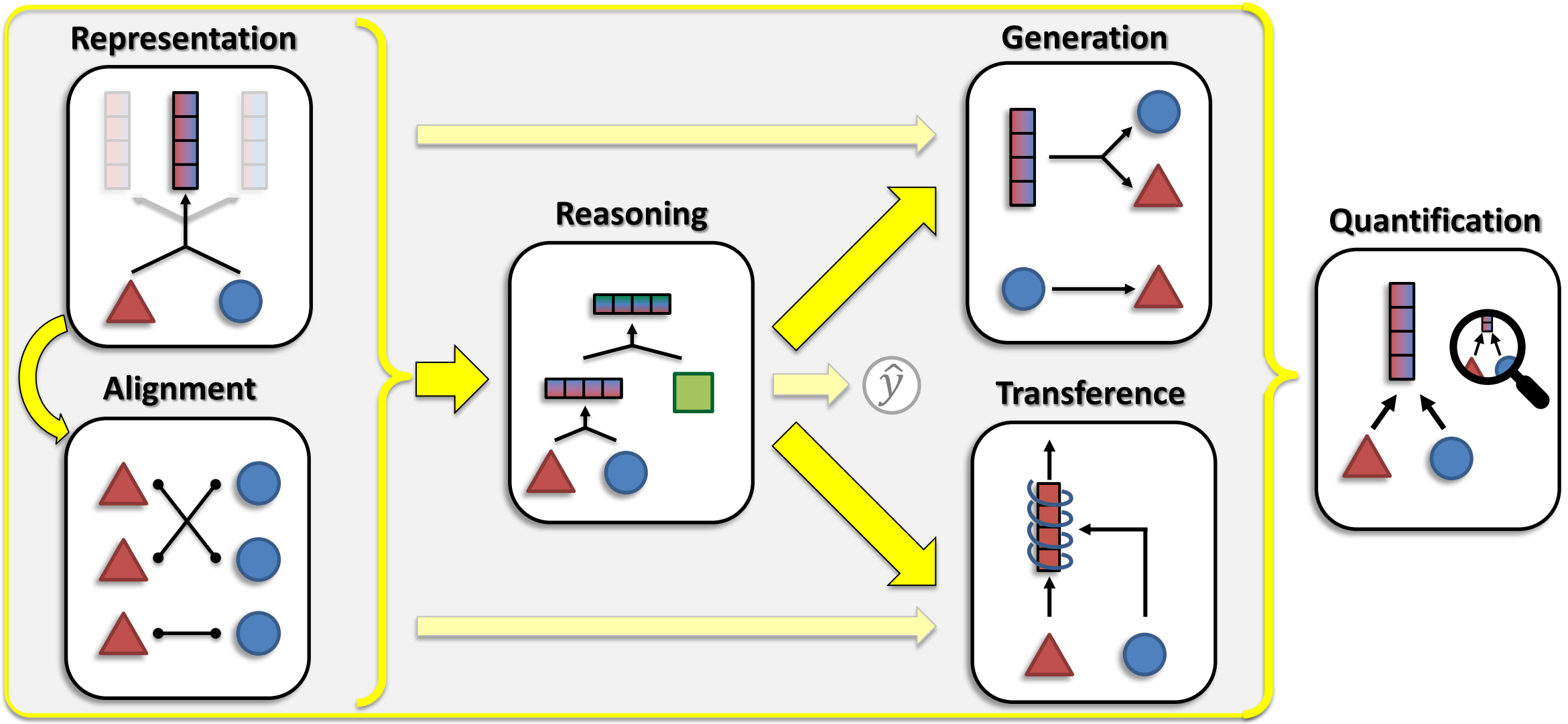


Learning



[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Core Multimodal Challenges



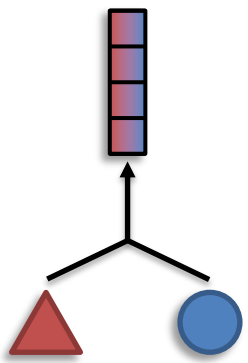
Challenge 1: Representation

Challenge 1: Representation

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

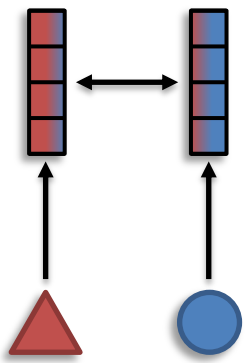
Sub-challenges:

Fusion



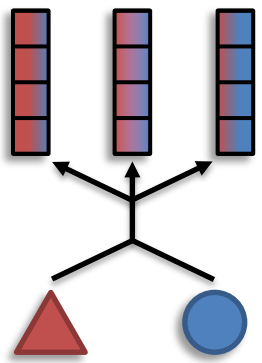
modalities **>** # representations

Coordination



modalities **=** # representations

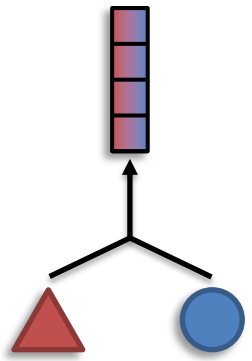
Fission



modalities **<** # representations

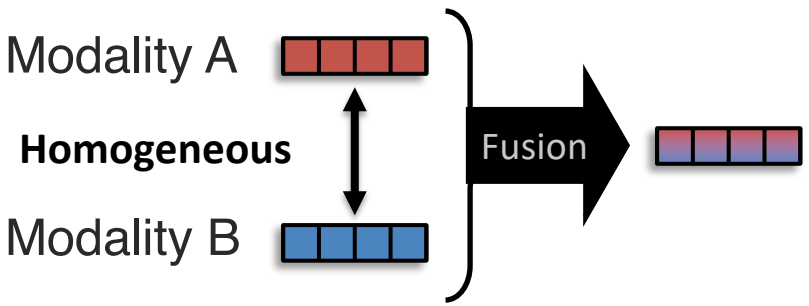
[Liang, Zadeh, and Morency. Foundations and Recent Trends on Multimodal Machine Learning. arXiv 2022]

Sub-Challenge 1a: Representation Fusion

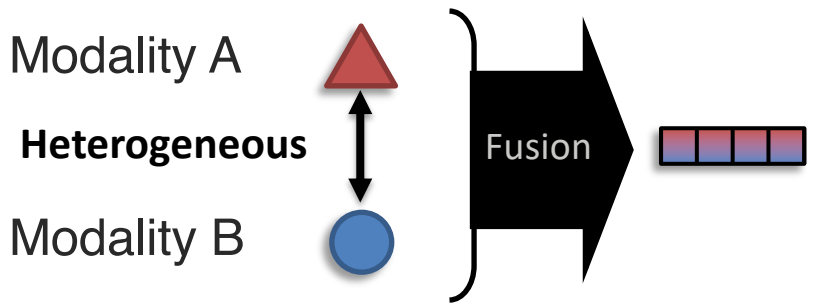


Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities.

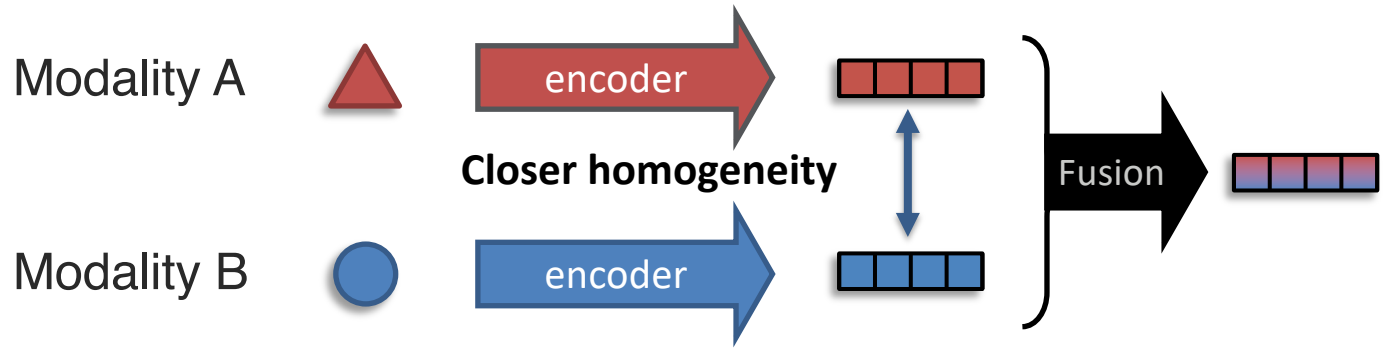
Fusion with abstract modalities:



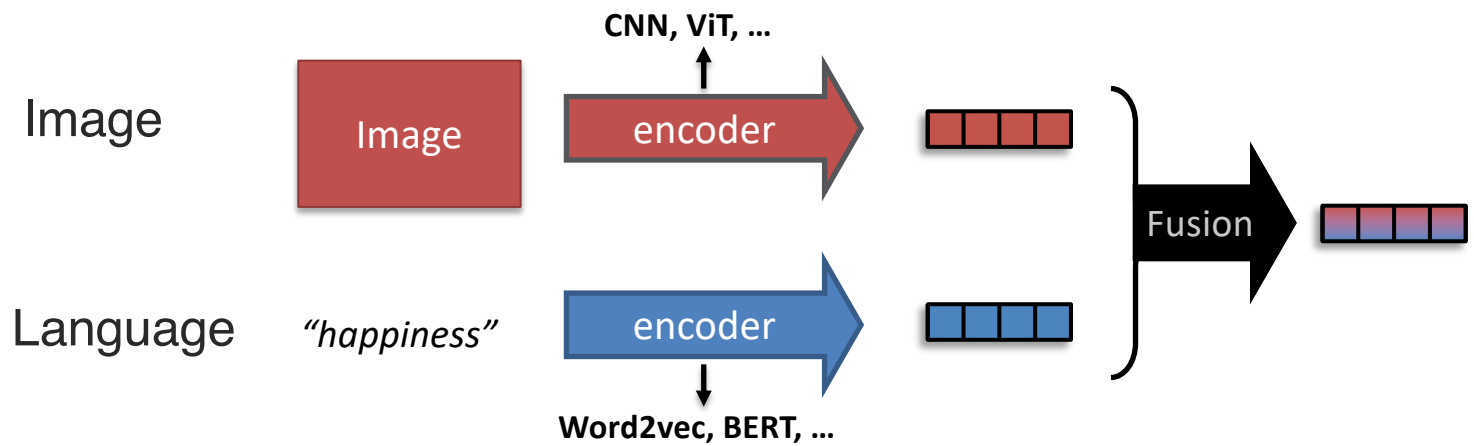
Fusion with raw modalities:



Fusion with Abstract Modalities

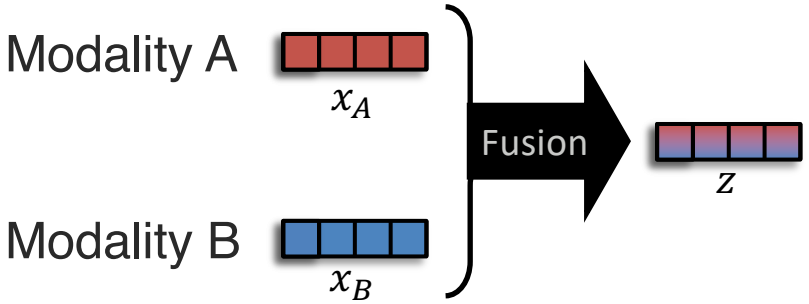


Example:



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

Basic Concepts for Representation Fusion



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

➡ Let's study the univariate case first

↳ (only 1-dimensional features)

Linear regression:

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

w_0: constant; $w_1 x_A + w_2 x_B$: Additive terms; $w_3 (x_A \times x_b)$: Multiplicative term; ϵ: error

① Additive interaction:

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

② Multiplicative interaction:

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

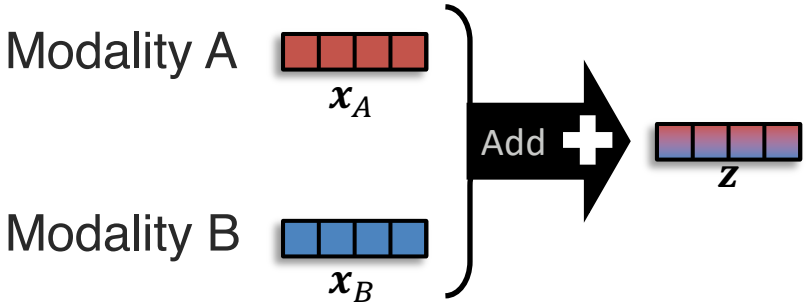
③ Additive and multiplicative interactions:

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

Additive Fusion

Back to multivariate case!

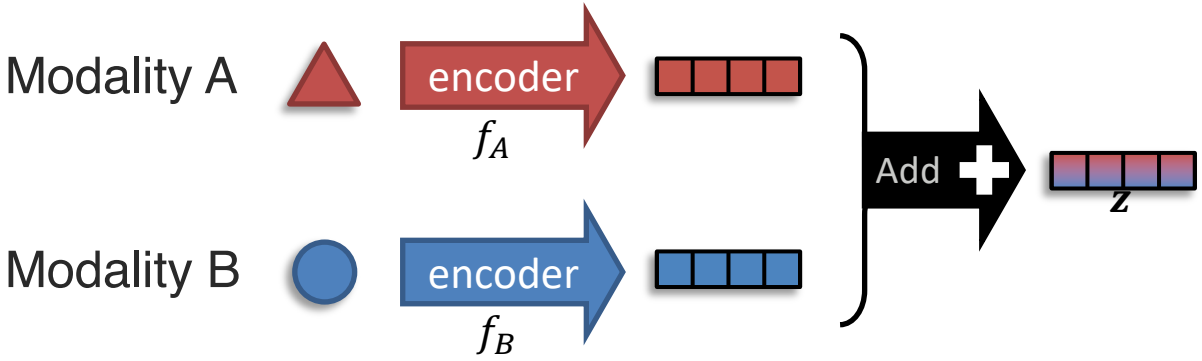
(multi-dimensional features)



Additive fusion:

$$z = w_1 x_A + w_2 x_B$$

With unimodal encoders:

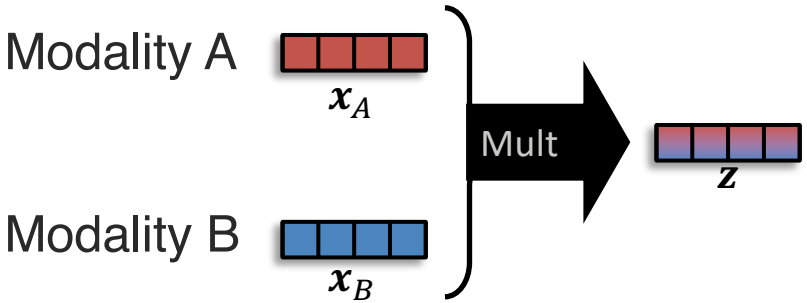


Additive fusion:

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

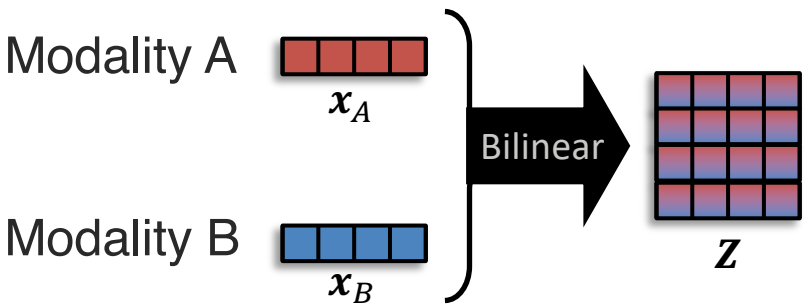
It could be seen as an ensemble approach (late fusion)

Multiplicative Fusion



Multiplicative fusion:

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

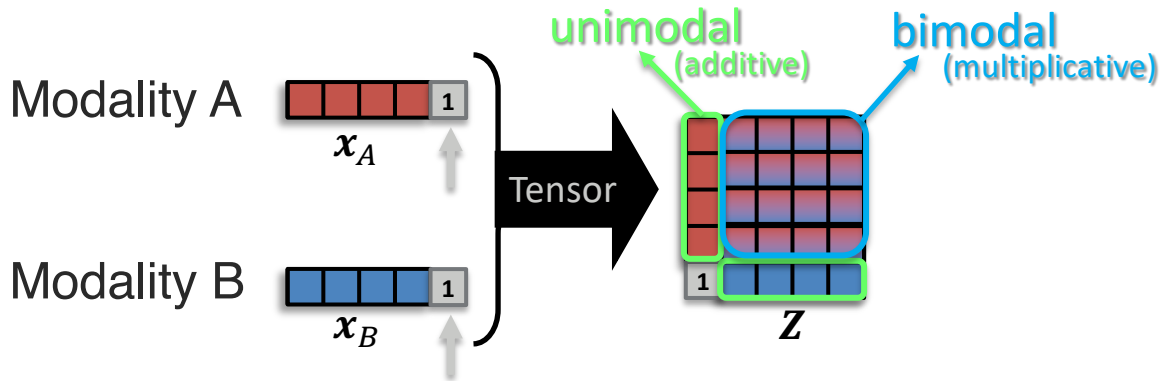


Bilinear Fusion:

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

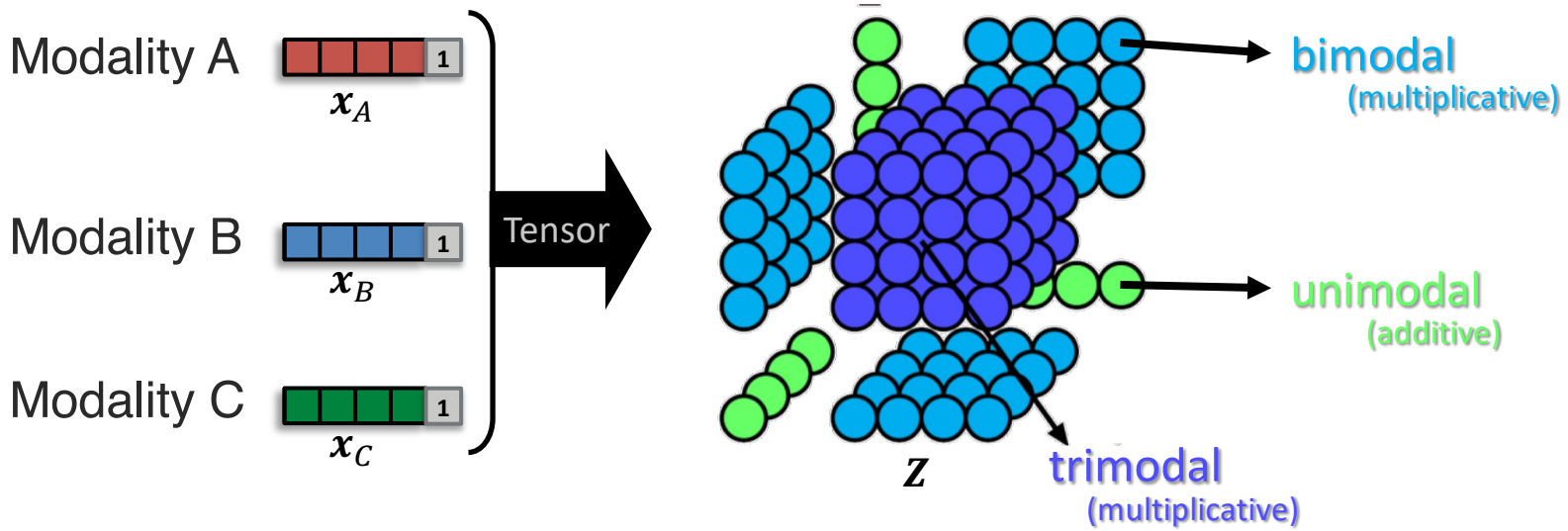
[Fukui et al., Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. EMNLP 2016]

Tensor Fusion



Tensor Fusion (bimodal):

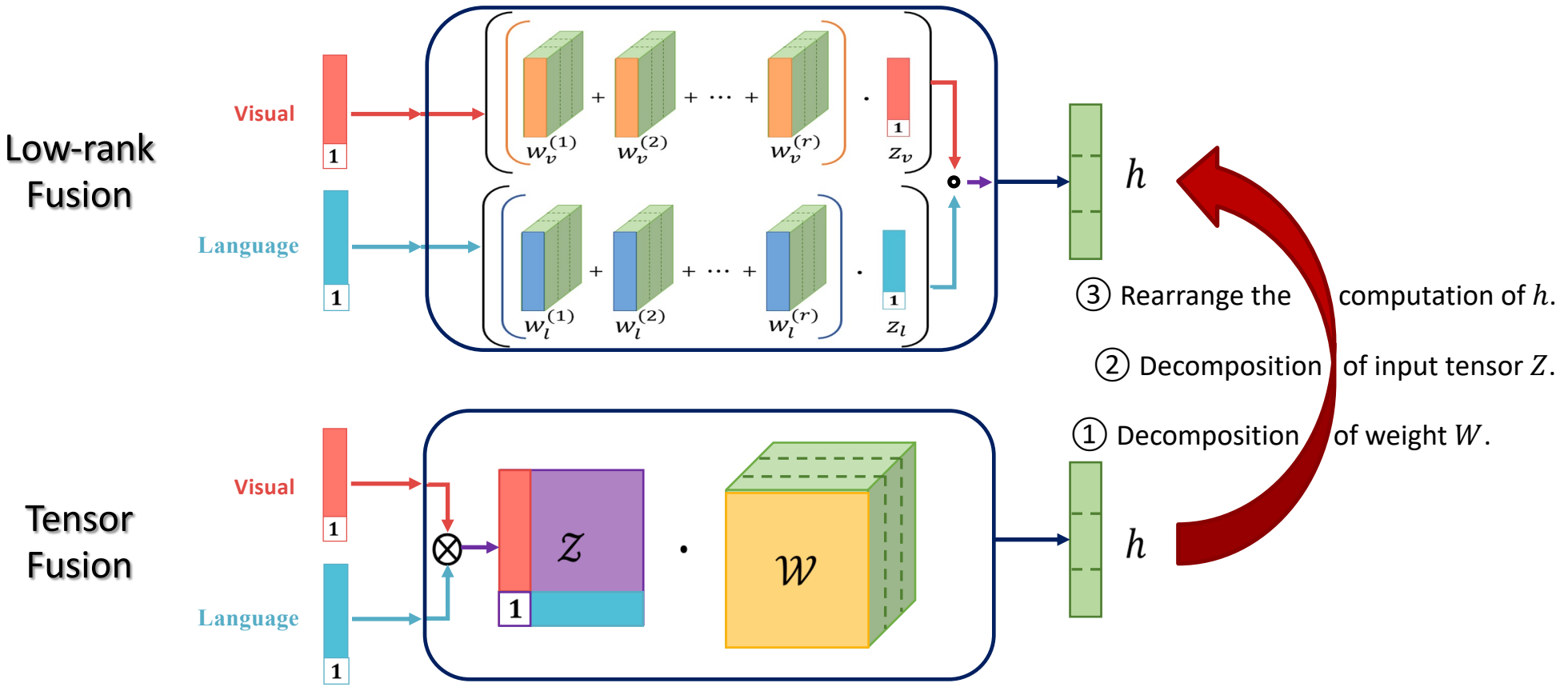
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

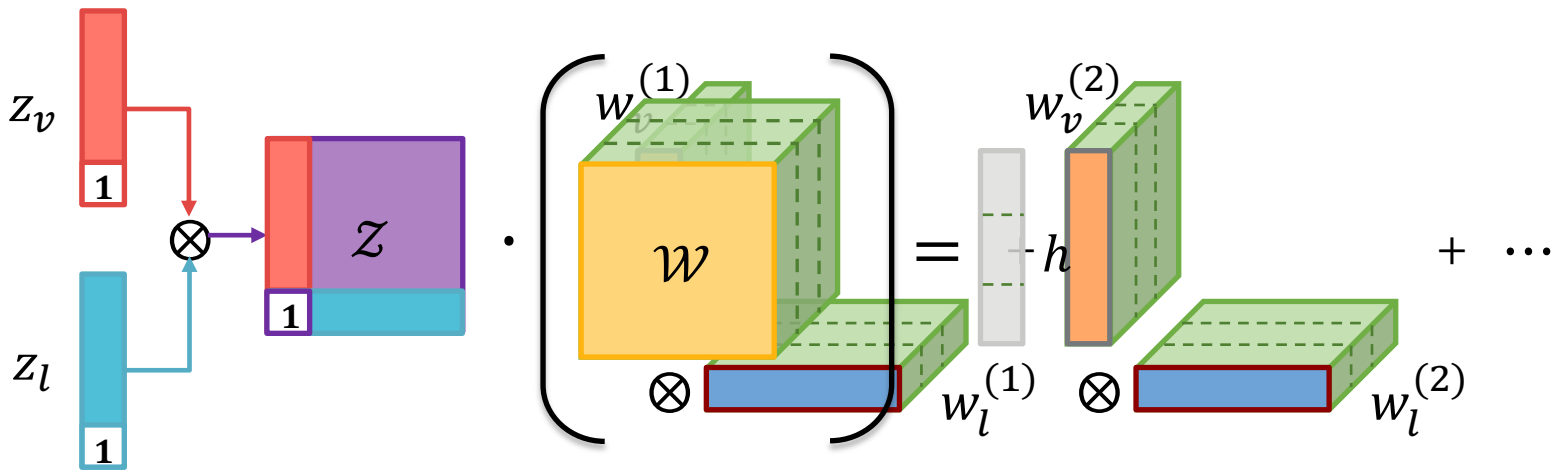
[Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017]

Low-rank Fusion



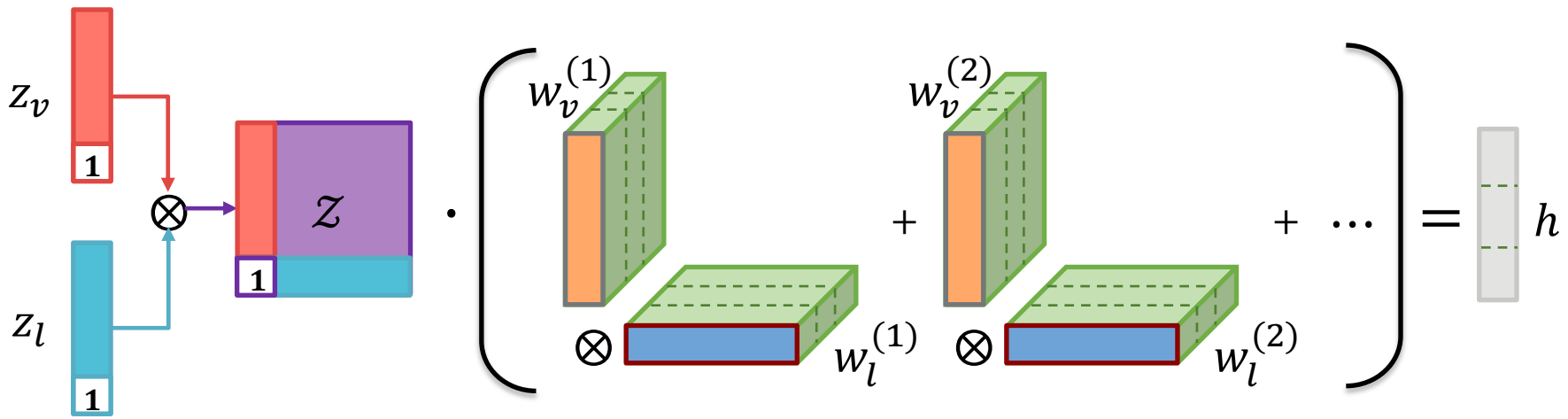
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Fusion



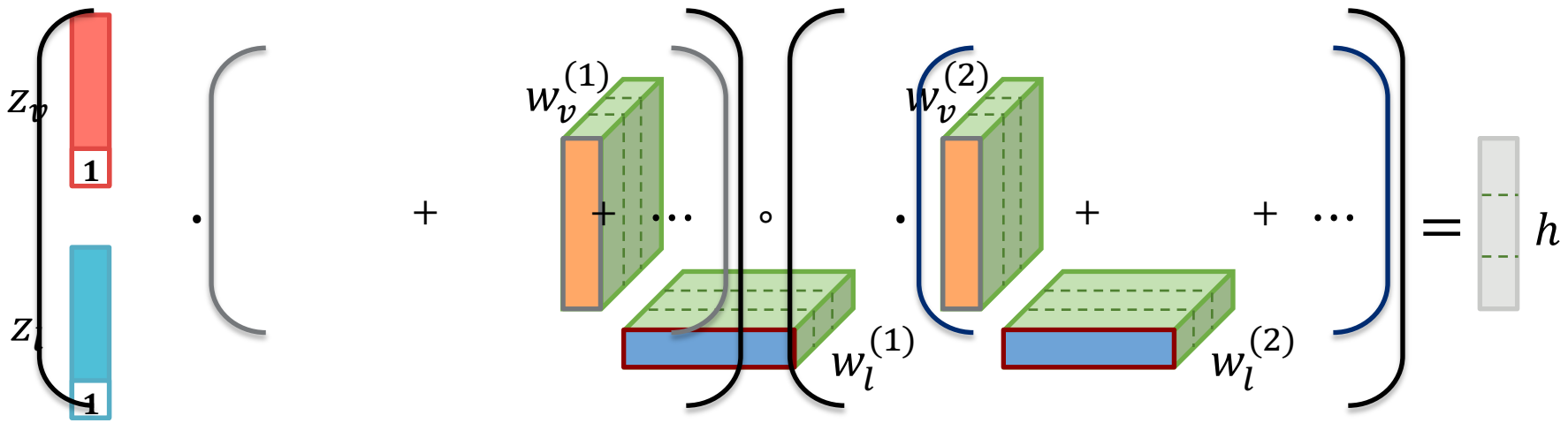
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]

Low-rank Fusion



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

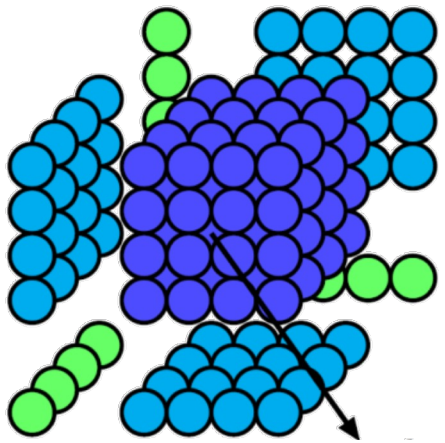
Low-rank Fusion



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

Low-rank Fusion with Trimodal Input

Tensor Fusion



Low-rank Fusion :

$$\left(\text{Low-rank factors}\ w_v^{(1)} + w_v^{(2)} + \cdots + w_v^{(r)}\right) \cdot z_v \circ \left(\text{Low-rank factors}\ w_l^{(1)} + w_l^{(2)} + \cdots + w_l^{(r)}\right) \cdot z_l \circ \left(\text{Low-rank factors}\ w_a^{(1)} + w_a^{(2)} + \cdots + w_a^{(r)}\right) \cdot z_a$$

Canonical Polyadic Decomposition

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

Going Beyond Additive and Multiplicative Fusion

Additive interaction:

$$z = w_1 x_A + w_2 x_B$$ ⟸ First-order polynomial

Additive and multiplicative interaction:

$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_B)$$ ⟸ Second-order polynomial

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

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

Can we add higher-order interaction terms?

For example:

$$+w_8 \left(x_A^2 \times x_b^2 \times x_c^2\right)$$

$$+w_9 \left(x_A^3 \times x_b\right)$$

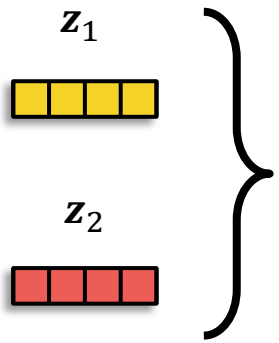
$$+w_{10} \left(x_b^3 \times x_c^3\right)$$

High-Order Polynomial Fusion

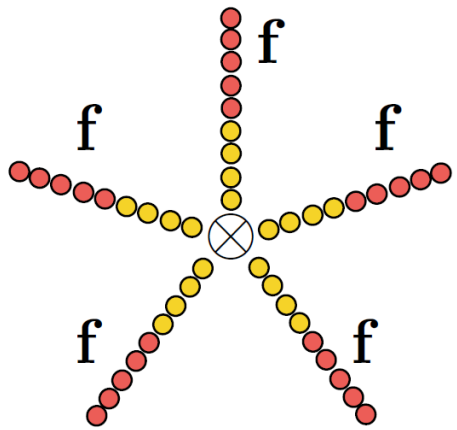
Concatenate

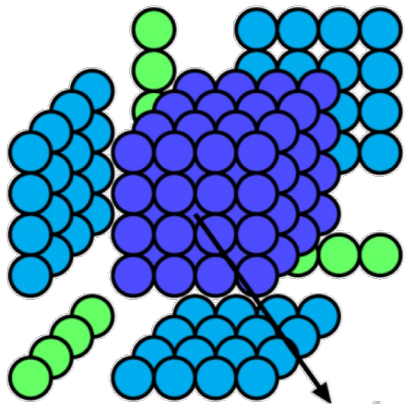
P-order tensor product

Weight matrix will be order P+1

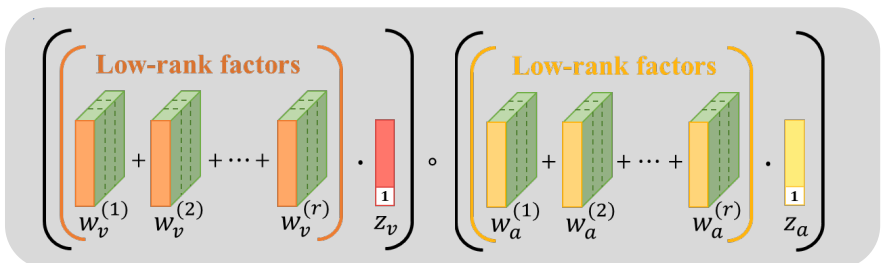


$$\mathbf{f}^\top = [1, \mathbf{z}_1^\top, \mathbf{z}_2^\top]$$

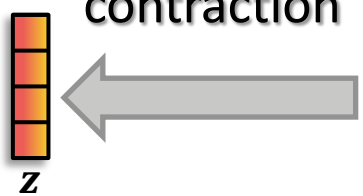




Low-rank tensor network

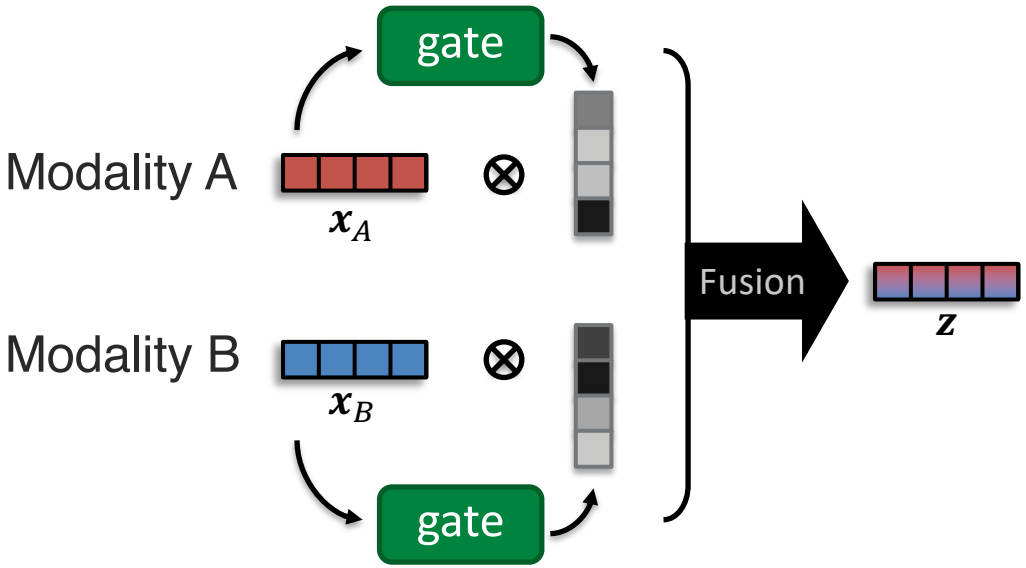


Tensor contraction



[Hou et al., Deep Multimodal Multilinear Fusion with High-order Polynomial Pooling. NeurIPS 2019]

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

Linear: $x_A W_1 \cdot (x_B W_2)^T$

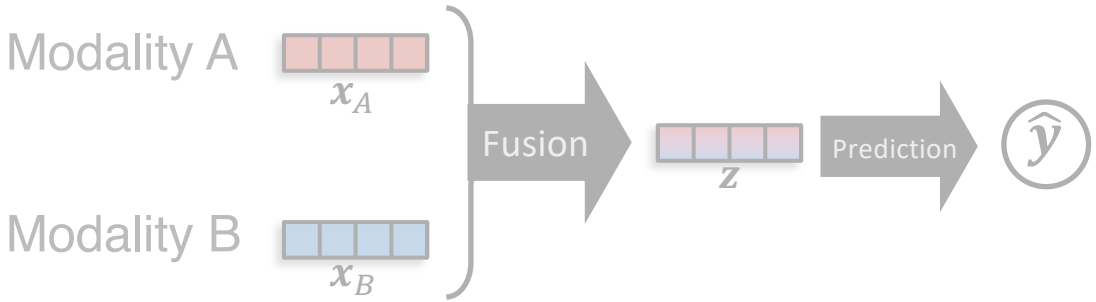
Nonlinear: $f_A(x_A) \cdot (f_B(x_B))^T$

Kernel: $k(x_A, x_B)$

- Linear
- Polynomial
- Exponential
- RBF
- Transformers

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

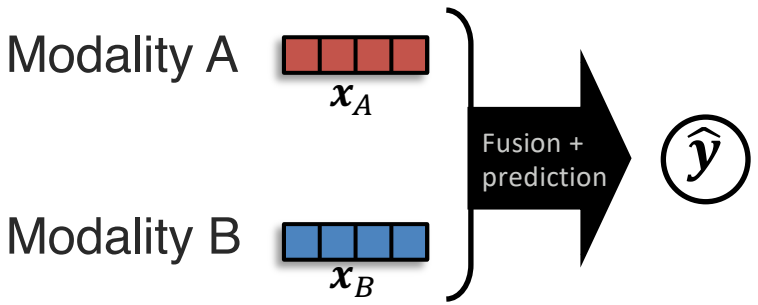
Nonlinear Fusion



Nonlinear fusion:

$$\hat{y} = f(x_A, x_B) \in \mathbb{R}^d$$

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



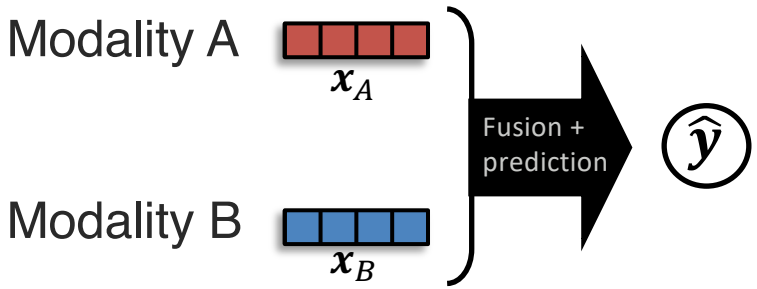
➡ This could be seen as *early fusion*:

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

… but will our neural network learn the nonlinear interactions?

More in the Quantification challenge!

Measuring Non-Additive Interactions



Nonlinear fusion:

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

Additive fusion:

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

Projection?

Projection from nonlinear to additive (using EMAP):

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

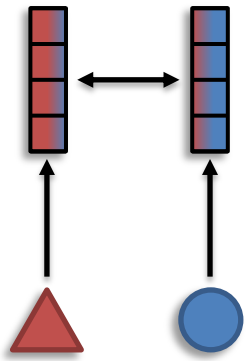
The expectations $\mathbb{E}$ can be approximated:

$$\hat{f}_A(x_A) = \frac{1}{N}\sum_{j=1}^{N} f(x_{A,j}, x_{B,j})$$

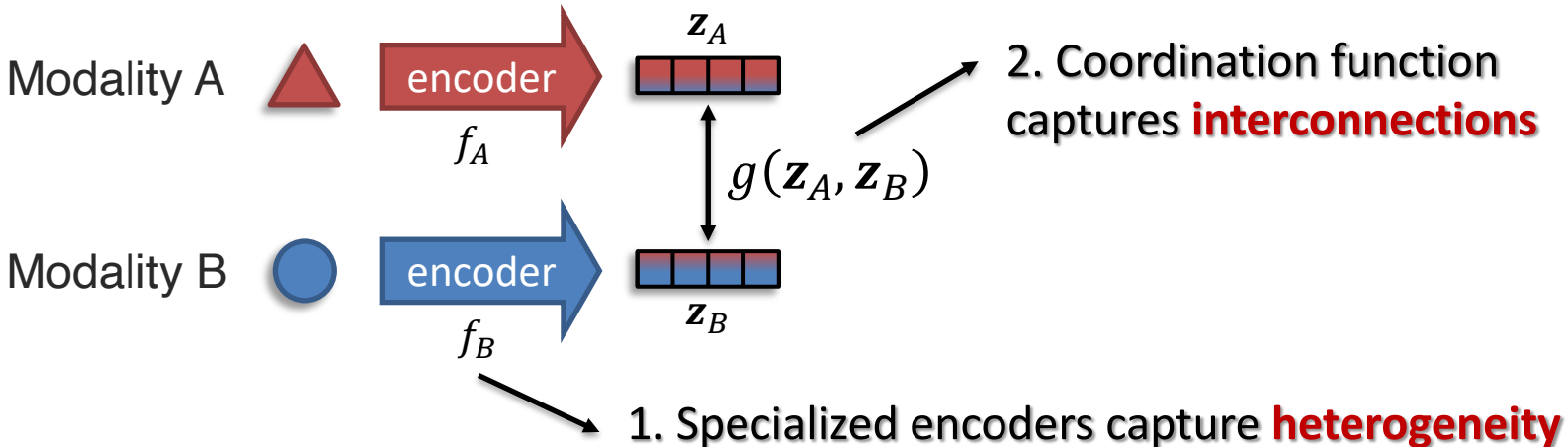
If the additive projection $\tilde{f}(x_A, x_B)$ is equal to nonlinear fusion $f(x_A, x_B)$ then the non-additive interactions are not properly modeled

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

Sub-Challenge 1b: Representation Coordination



Definition: Learn multimodal contextualized representations coordinated through their interconnections.

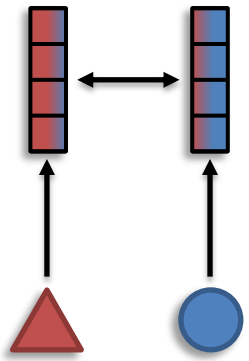


Learning with coordination function:

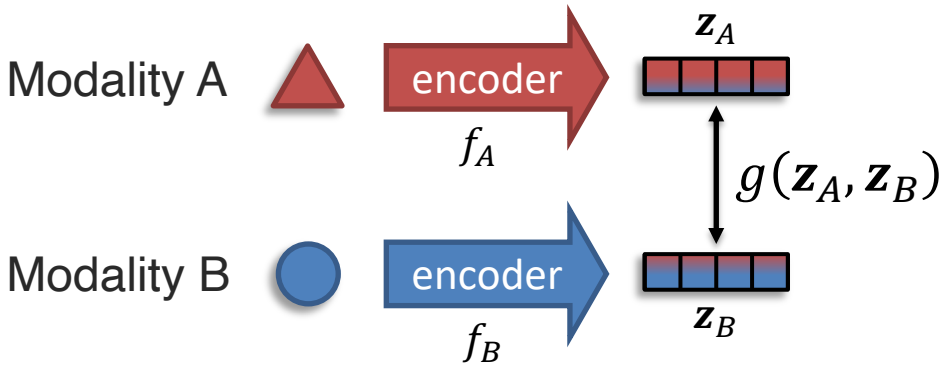
$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

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

Coordinated Representations



Definition: Learn multimodal contextualized representations coordinated through their interconnections.



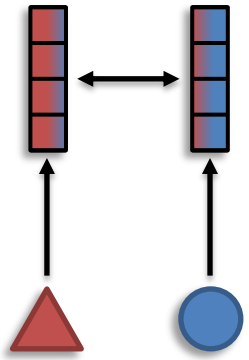
Learning with coordination function:

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

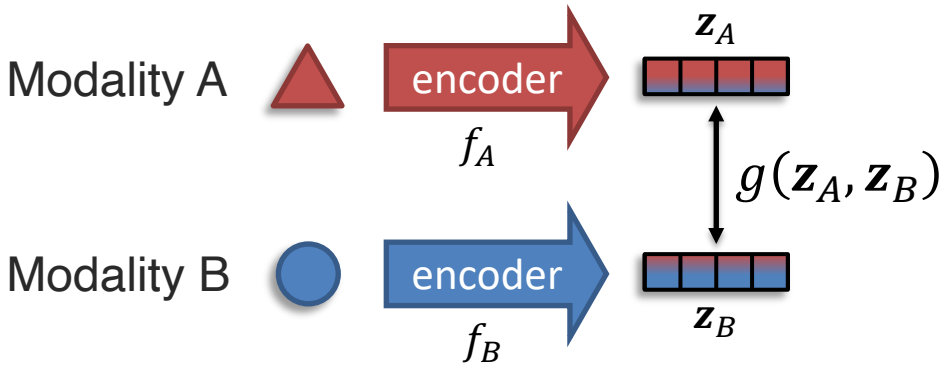
with model parameters θ_g, θ_{f_A} and θ_{f_B}

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

Coordinated Representations



Definition: Learn multimodal contextualized representations coordinated through their interconnections.



Learning with coordination function:

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

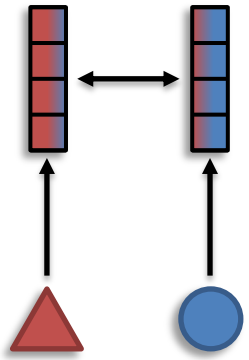
with model parameters θ_g, θ_{f_A} and θ_{f_B}

② Kernel similarity functions:

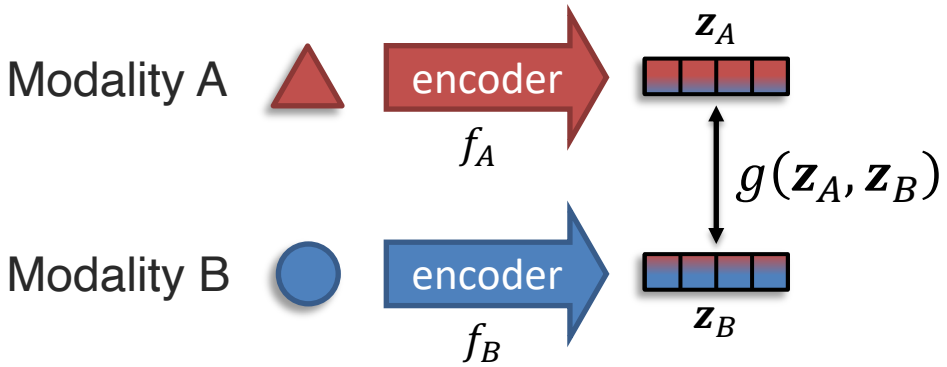
$$g(\boldsymbol{z}_A, \boldsymbol{z}_B) = k(\boldsymbol{z}_A, \boldsymbol{z}_B)$$

- Linear
- Polynomial
- Exponential
- RBF

Coordinated Representations



Definition: Learn multimodal contextualized representations coordinated through their interconnections.



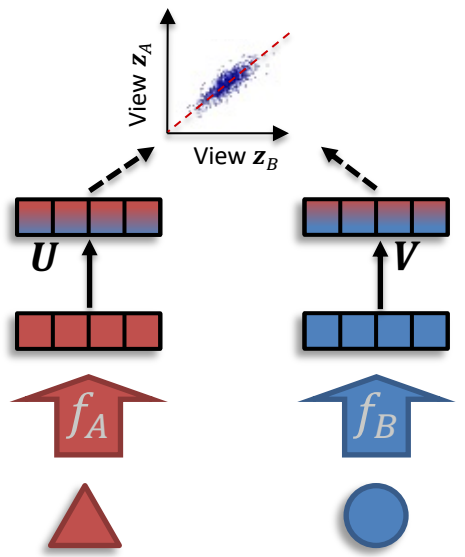
Learning with coordination function:

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

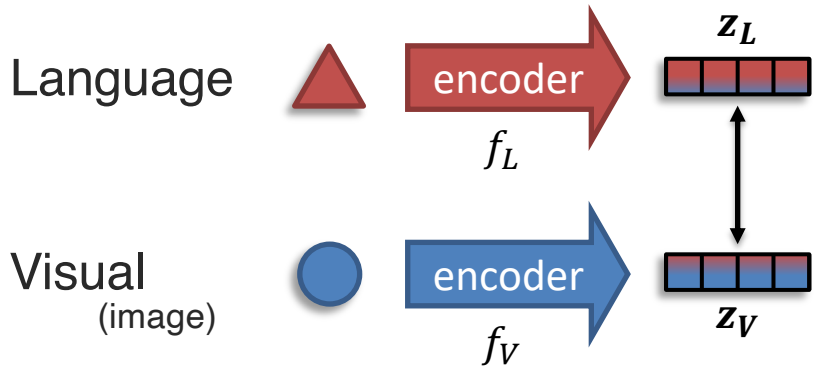
with model parameters θ_g, θ_{f_A} and θ_{f_B}

③ Canonical Correlation Analysis (CCA):

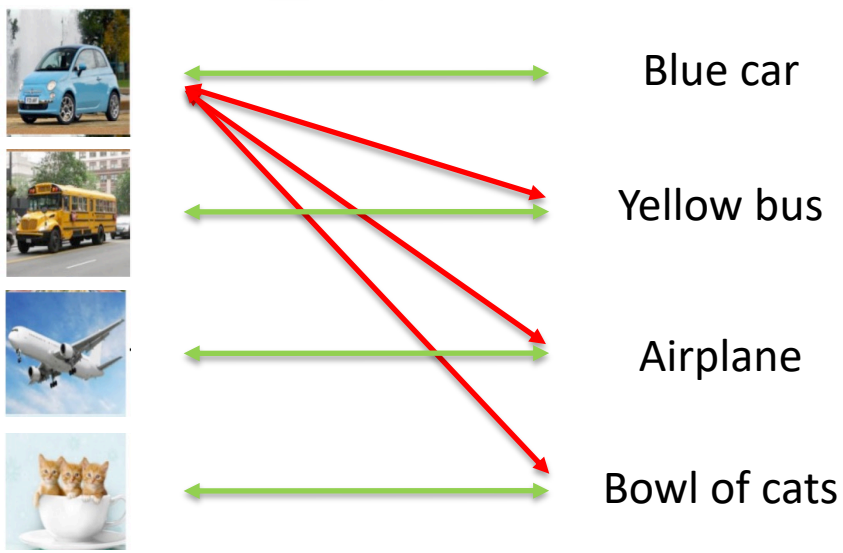
$$\underset{\boldsymbol{V},\boldsymbol{U},f_A,f_B}{\operatorname{argmax}}\ corr(\boldsymbol{z}_A, \boldsymbol{z}_B)$$



Coordination with Contrastive Learning



Paired data: {▲, ●} (e.g., images and text)



Contrastive loss:

⇨ brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

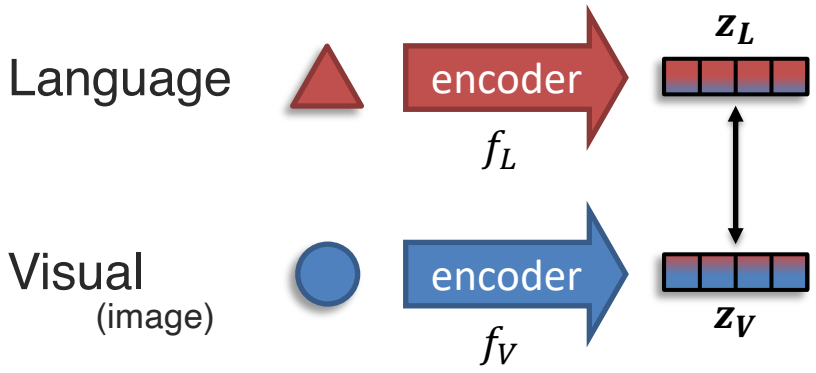
$$\max\{0, \alpha + g(\boldsymbol{z}_A, \boldsymbol{z}_B^+) - g(\boldsymbol{z}_A, \boldsymbol{z}_B^-)\}$$

positive pairs

negative pair

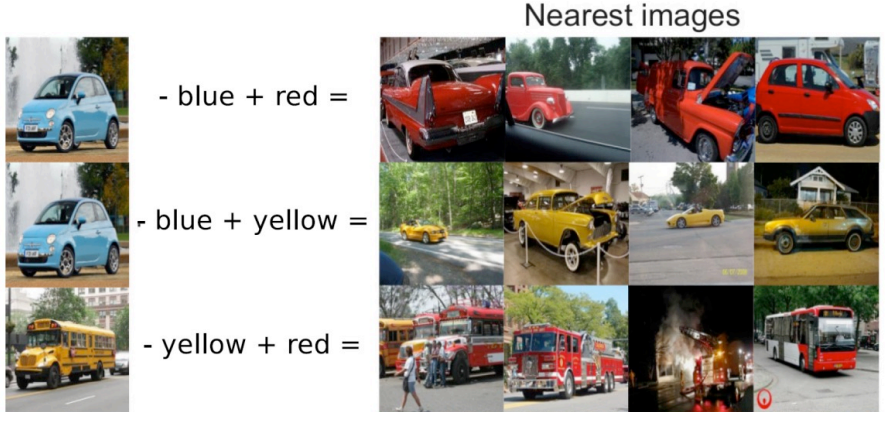
Coordination function
(e.g., cosine similarity)

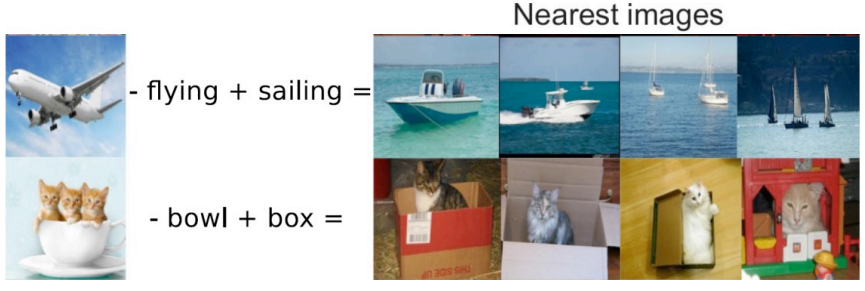
Visual-Semantic Embeddings



Contrastive loss:

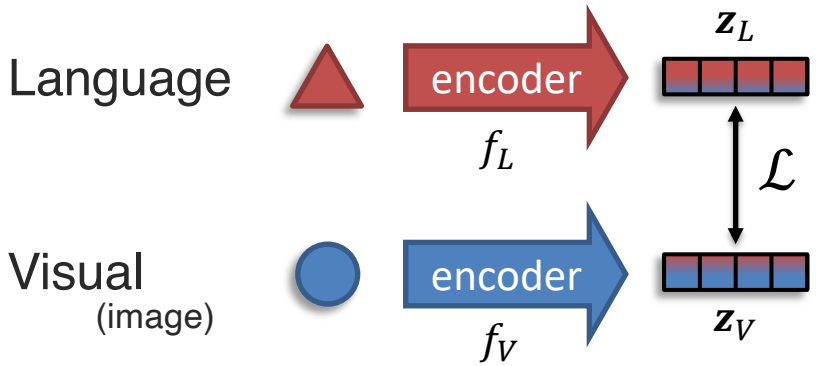
- brings positive pairs closer and pushes negative pairs apart





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

CLIP (Contrastive Language–Image Pre-training)



Popular contrastive loss: InfoNCE

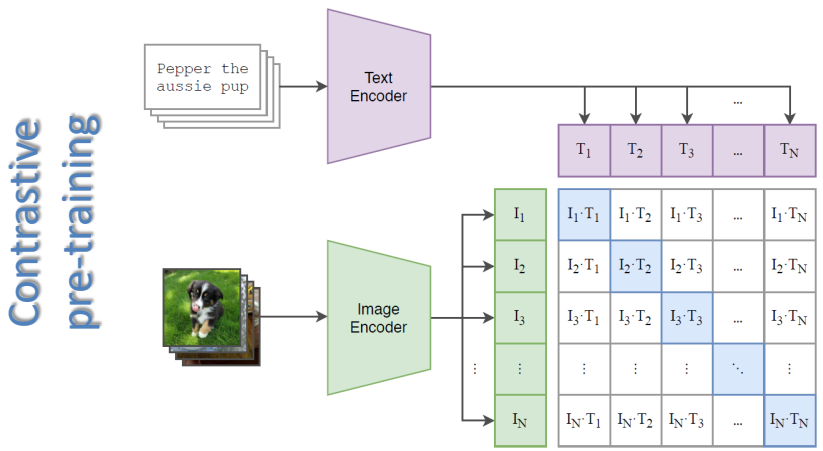
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{sim(\boldsymbol{z}_A^i, \boldsymbol{z}_B^i)}{\sum_{j=1}^{N} sim(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)}$$

positive pairs

negative pairs and positive pairs

Similarity function can be cosine similarity

Positive and negative pairs:

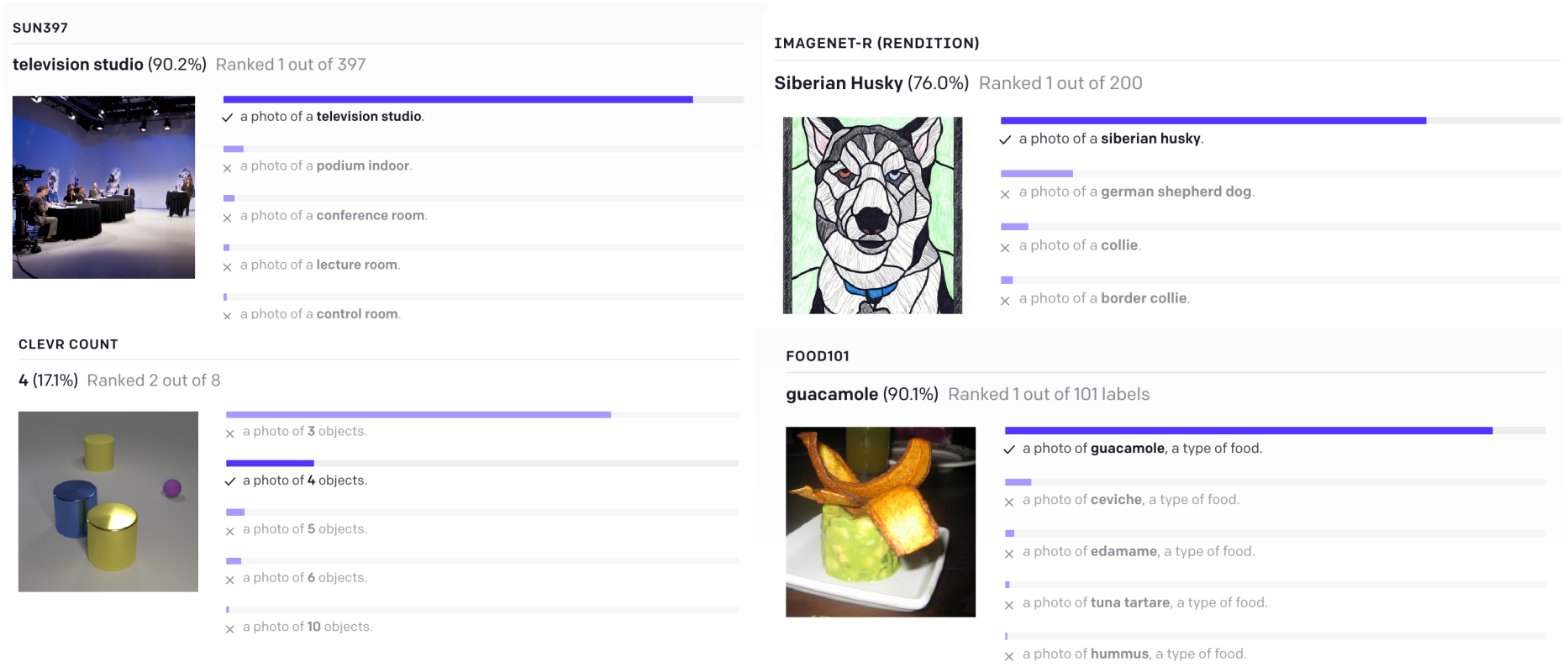


→ f_L and f_V are great encoders for language-vision tasks

→ z_L and z_V are coordinated but not identical representation spaces

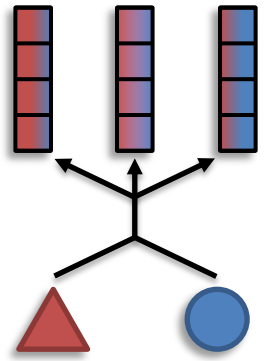
[Ramesh et al., Learning Transferable Visual Models from Natural Language Supervision. ICML 2022]

CLIP (Contrastive Language–Image Pre-training)



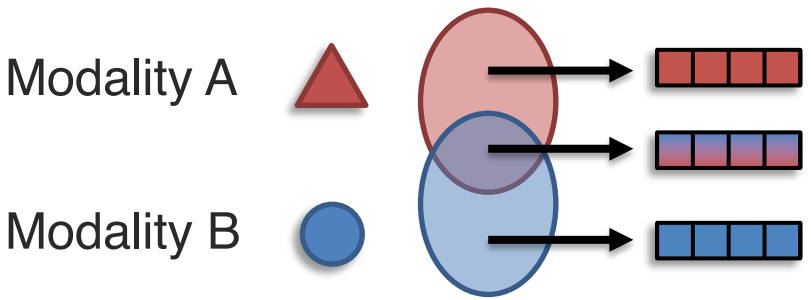
[Ramesh et al., Learning Transferable Visual Models from Natural Language Supervision. ICML 2022]

Sub-Challenge 1c: Representation Fission

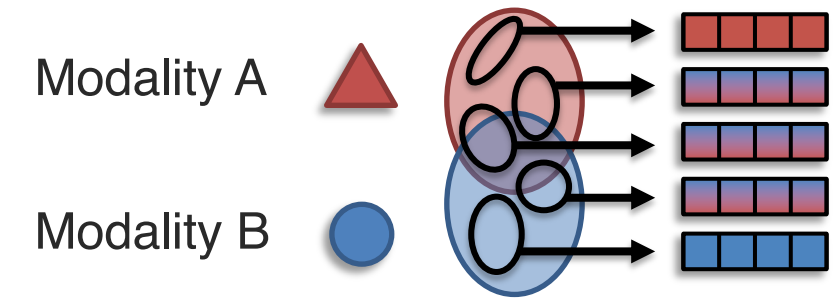


Definition: Learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

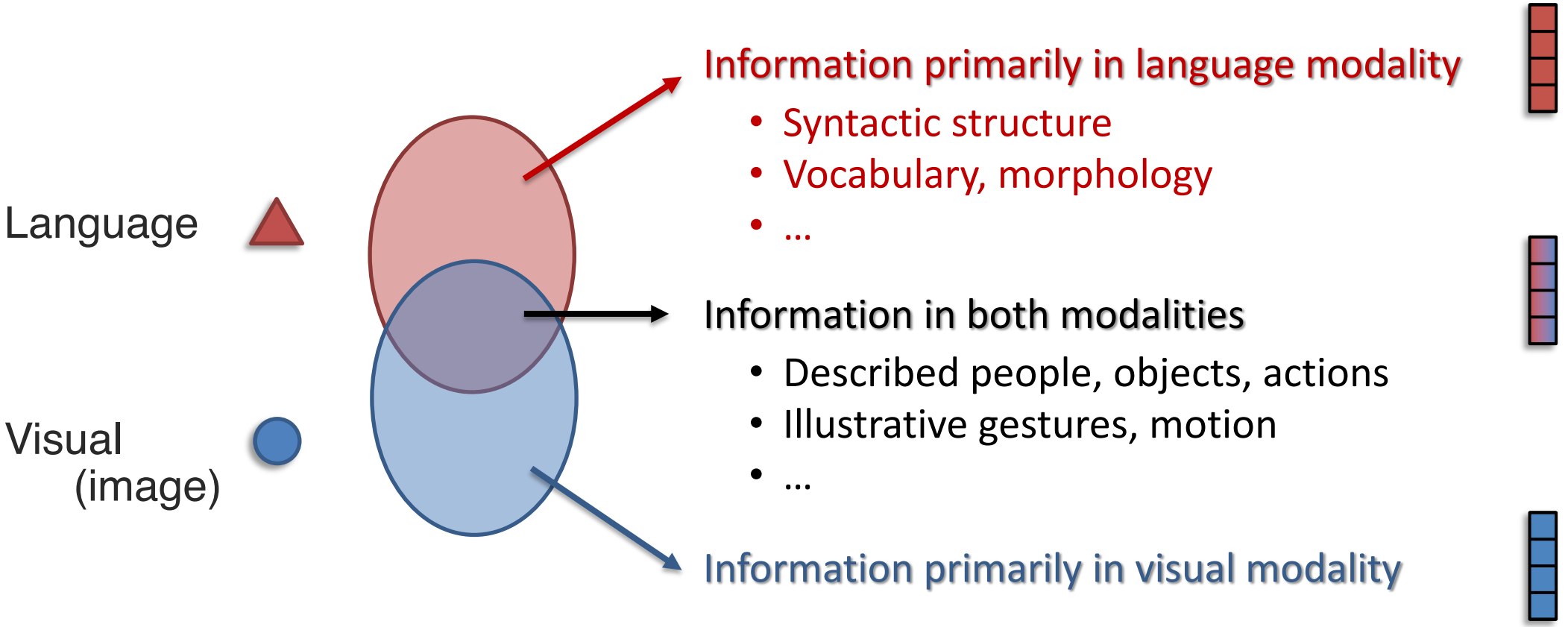
Modality-level fission:



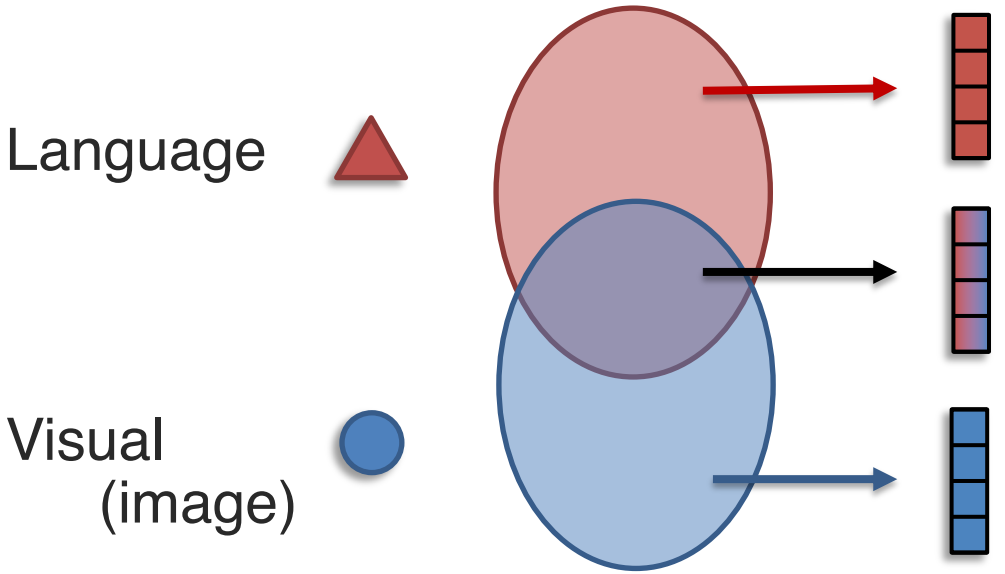
Fine-grained fission:



Modality-Level Fission

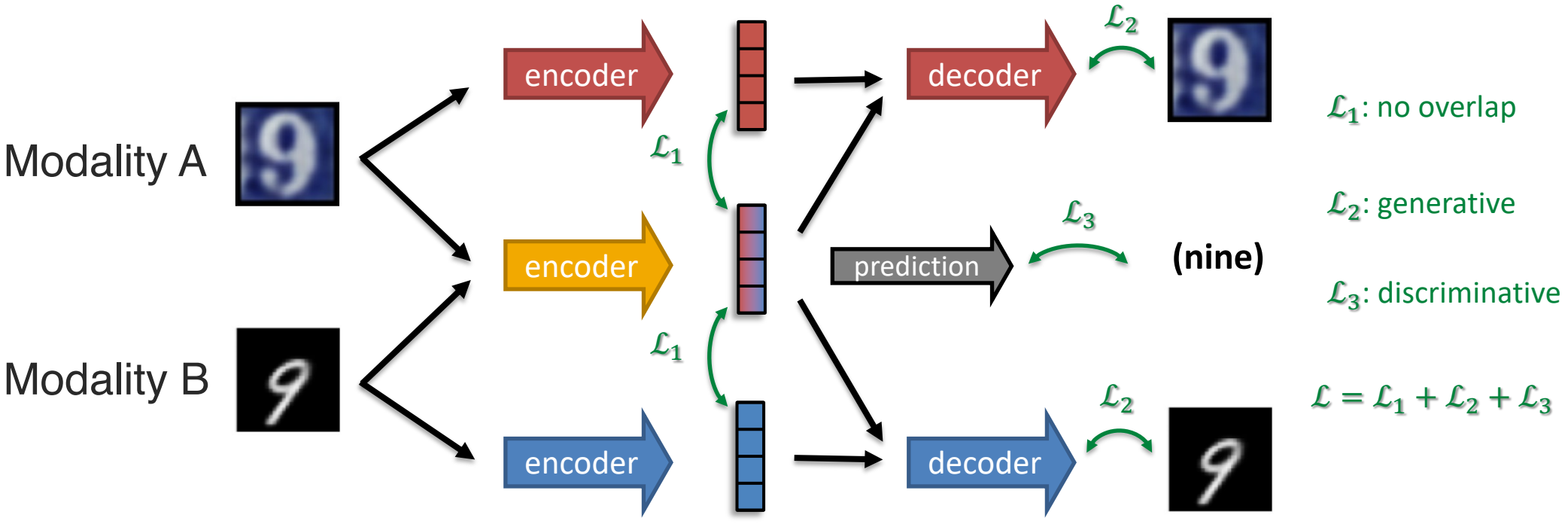


Modality-Level Fission



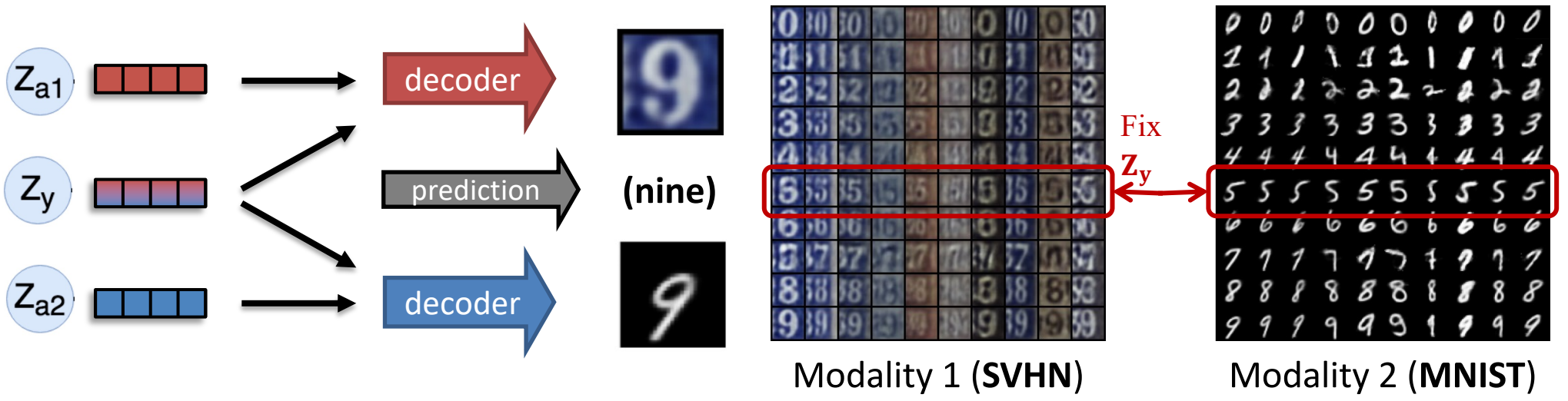
How to learn these disentangled representations?

Learning Factorized Multimodal Representations



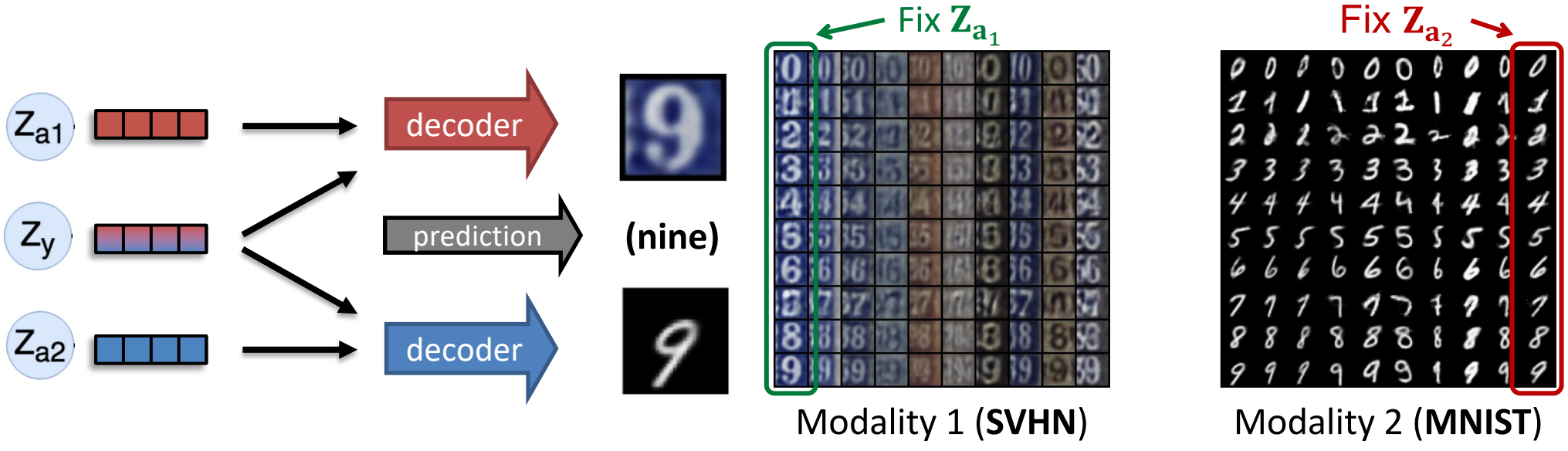
[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Learning Factorized Multimodal Representations



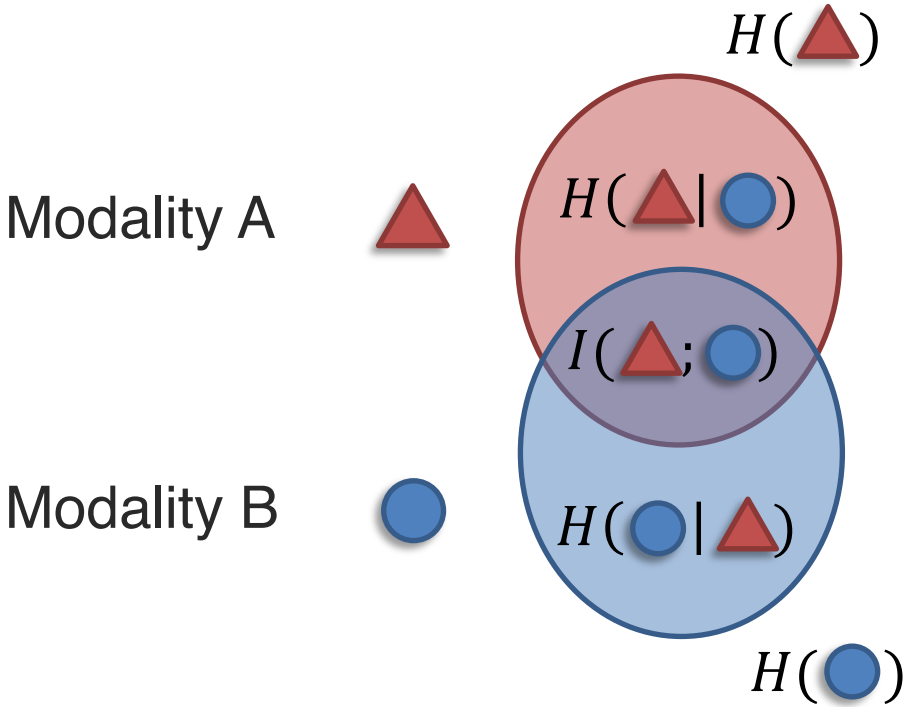
[Tsai et al., Learning Factorized Multimodal Representations. ICLR 2019]

Learning Factorized Multimodal Representations



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

Information Theory Perspective



Entropy:

$$H(X) := -\sum_{x \in \mathcal{X}} p(x) log_b p(x)$$

➡ measurement of degree of randomness, uncertainty

Mutual information:

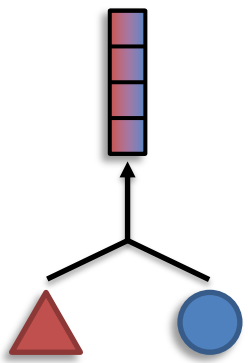
$$I(X;Y) = D_{KL}\left(P_{(X,Y)} \parallel P_X \otimes P_Y\right)$$

Challenge 1: Representation

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

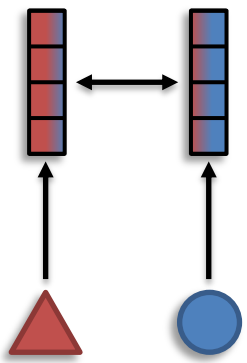
Sub-challenges:

Fusion



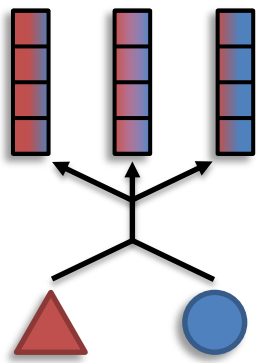
modalities **>** # representations

Coordination



modalities **=** # representations

Fission



modalities **<** # representations

[Liang, Zadeh, and Morency. Foundations and Recent Trends on Multimodal Machine Learning. arXiv 2022]

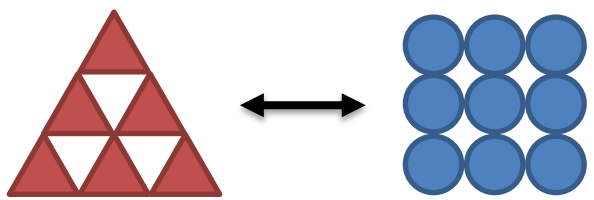
Challenge 6: Quantification

Challenge 6: Quantification

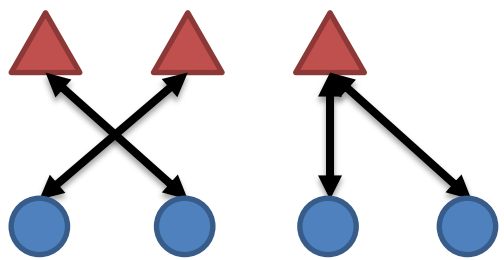
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions, and the multimodal learning process.

Sub-challenges:

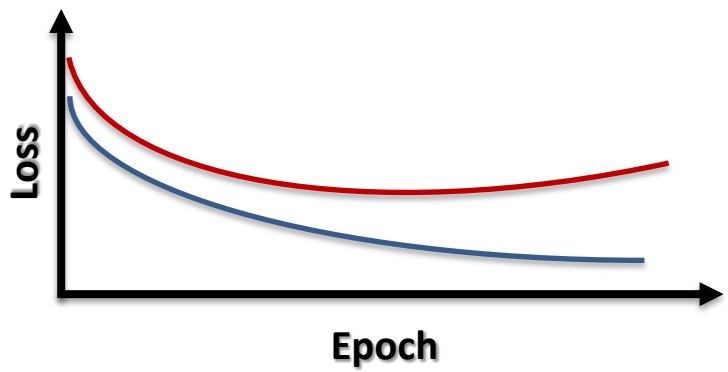
Heterogeneity



Interconnections



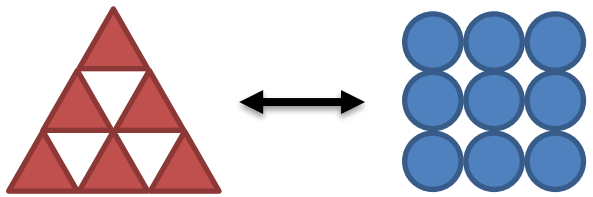
Learning



[Liang, Zadeh, and Morency. Foundations and Recent Trends in Multimodal Machine Learning. arXiv 2022]

Sub-Challenge 6a: Heterogeneity

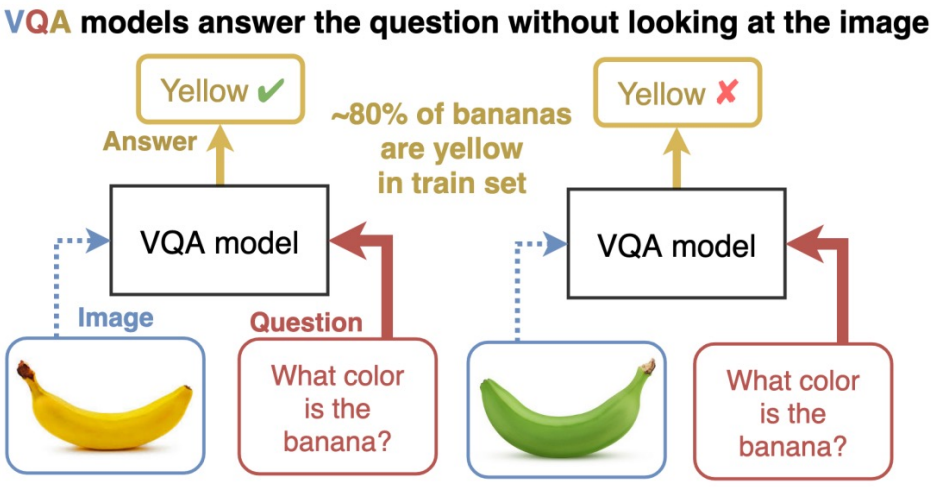
Definition: Quantifying the dimensions of heterogeneity in multimodal datasets and how they subsequently influence modeling and learning.



1. **Distribution:** discrete, continuous, support
2. **Granularity:** sampling rate, resolution, granularity
3. **Information:** entropy, density, information overlap, range
4. **Structure:** static, temporal, spatial, hierarchical, invariances
5. **Noise:** uncertainty, signal-to-noise ratio, missing data

Modality Biases

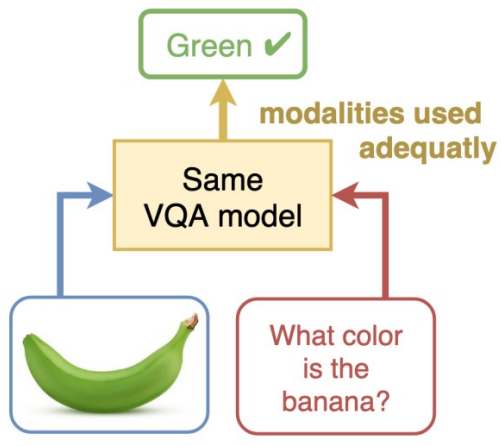
Unimodal biases and modality collapse
(Heterogeneity in information and relevance)



Balancing modalities

Balancing training

Not the case when trained with RUBi



[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]
[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]

Modality Biases

Fairness and social biases – unimodal social biases
(Heterogeneity in information and relevance)

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

Wrong

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

[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

Modality Biases

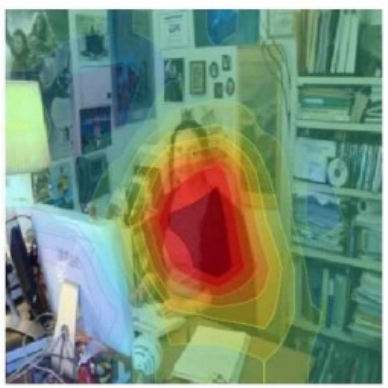
Fairness and social biases – unimodal social biases
(Heterogeneity in information and relevance)

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

Wrong

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

Right for the Right Reasons



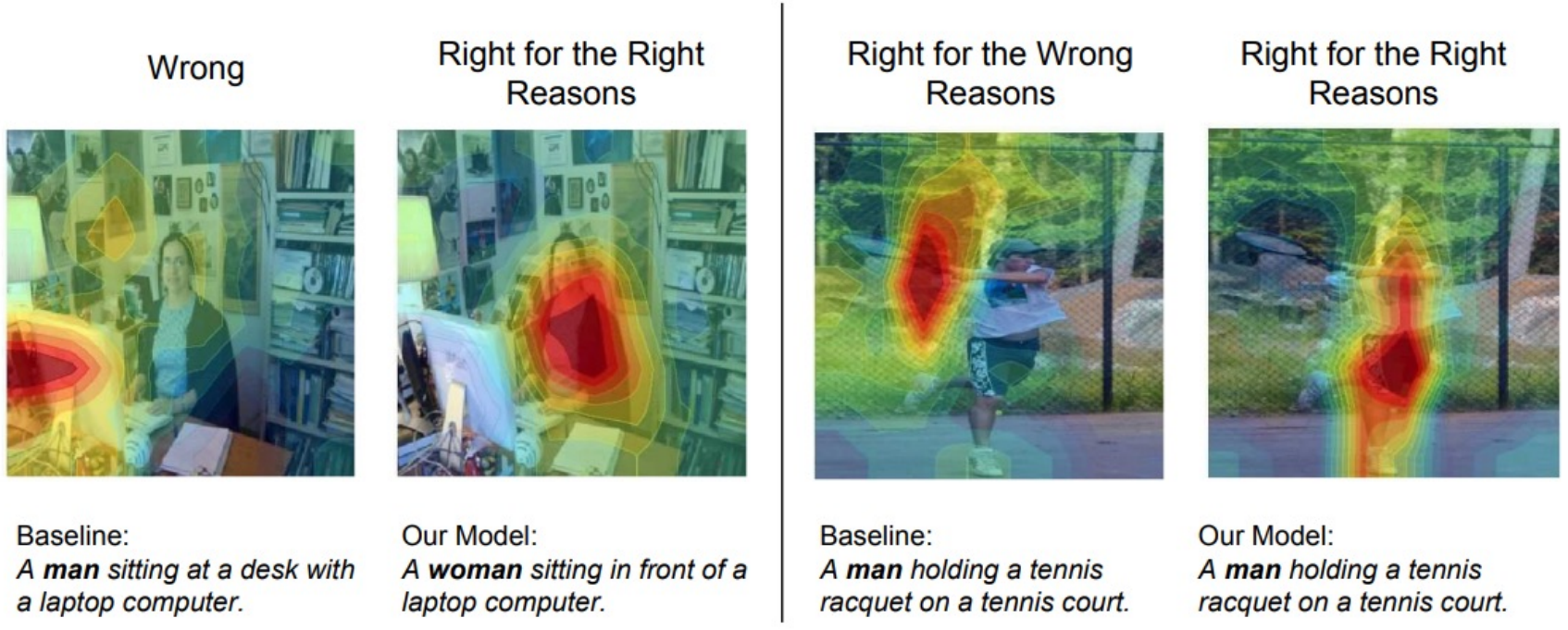
Our Model:
*A **woman** sitting in front of a laptop computer.*

[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

Modality Biases

Fairness and social biases – unimodal social biases
(Heterogeneity in information and relevance)

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

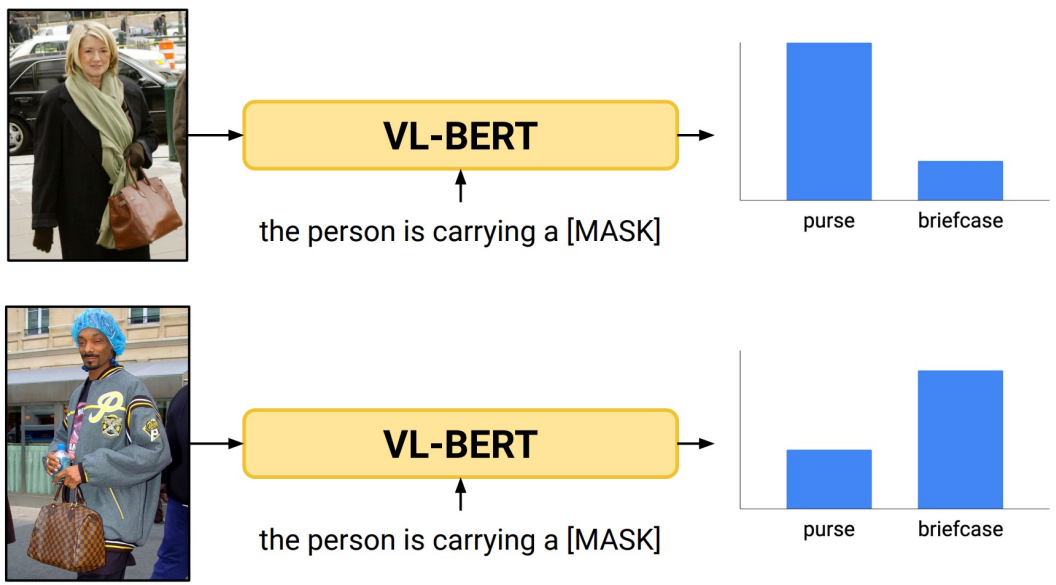


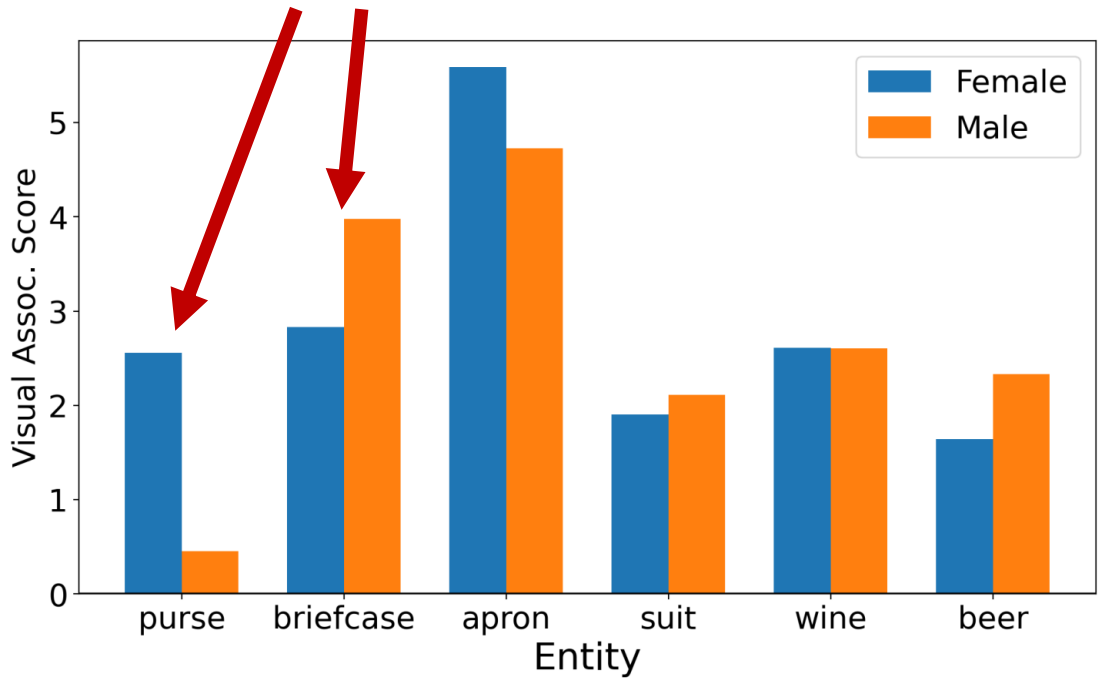
[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

Modality Biases

Fairness and social biases – cross-modal interactions worsen social biases
(Heterogeneity in information and relevance)

Visual information makes model more confident in reinforcing gender stereotypes





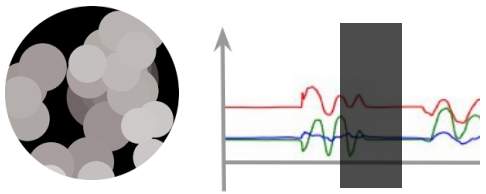
[Srinivasan and Bisk, Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. NAACL 2022]

Noise Topologies and Robustness

Heterogeneity in noise

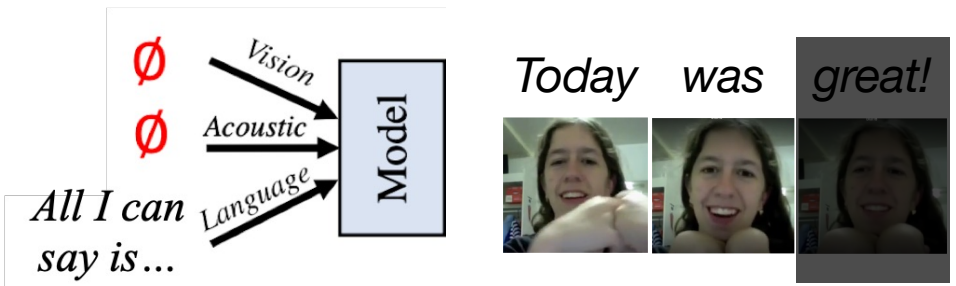
Modality-specific robustness

noise → nosie



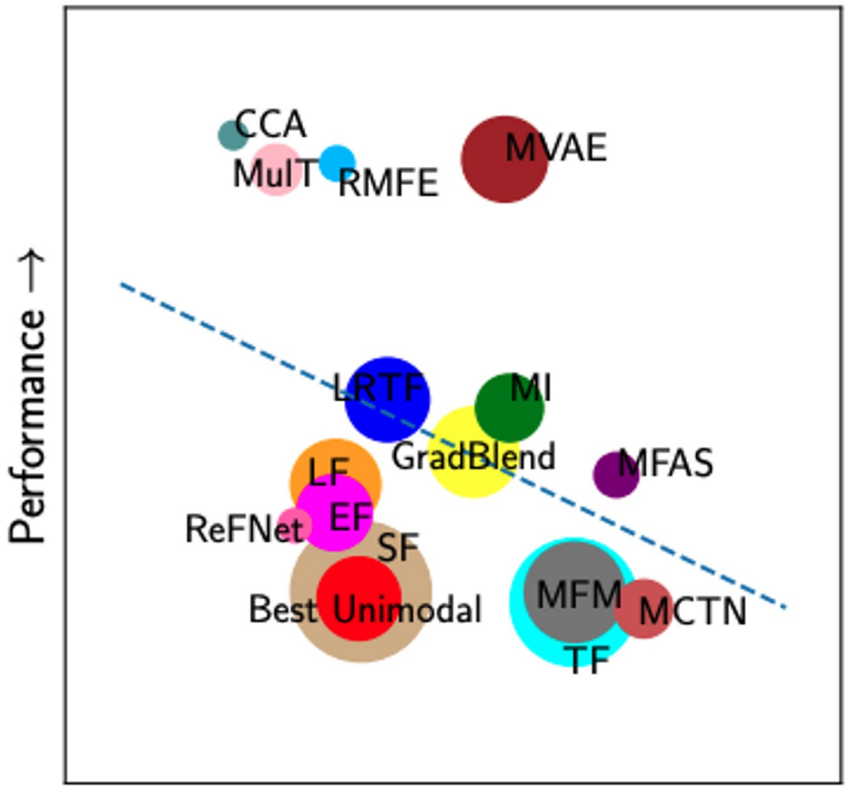
[Belinkov & Bisk, 2018; Subramaniam et al., 2009; Boyat & Joshi, 2015]

Multimodal robustness



[Zadeh et al., 2020]

Strong tradeoffs between performance and robustness



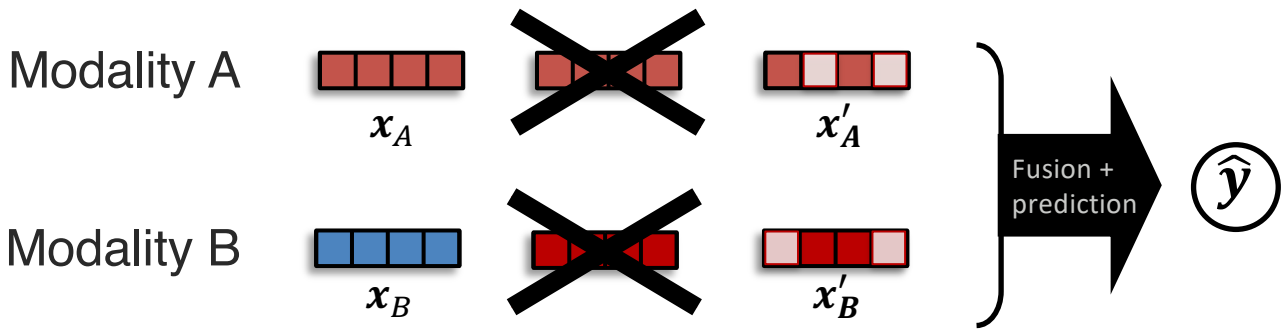
rate of accuracy drops

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

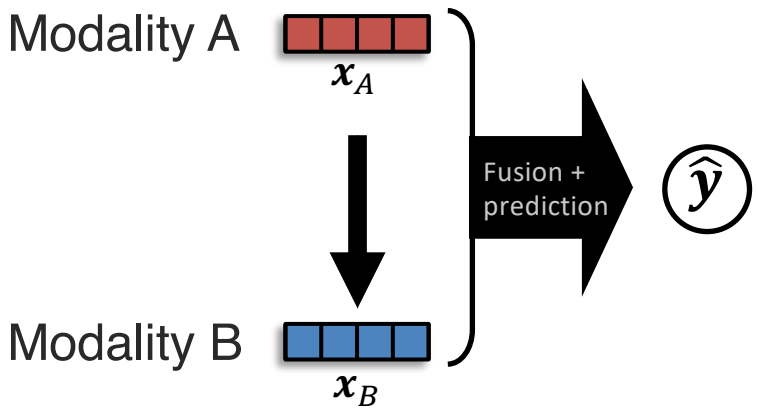
Noise Topologies and Robustness

Several approaches towards more robust models

Robust data + training



Infer missing modalities

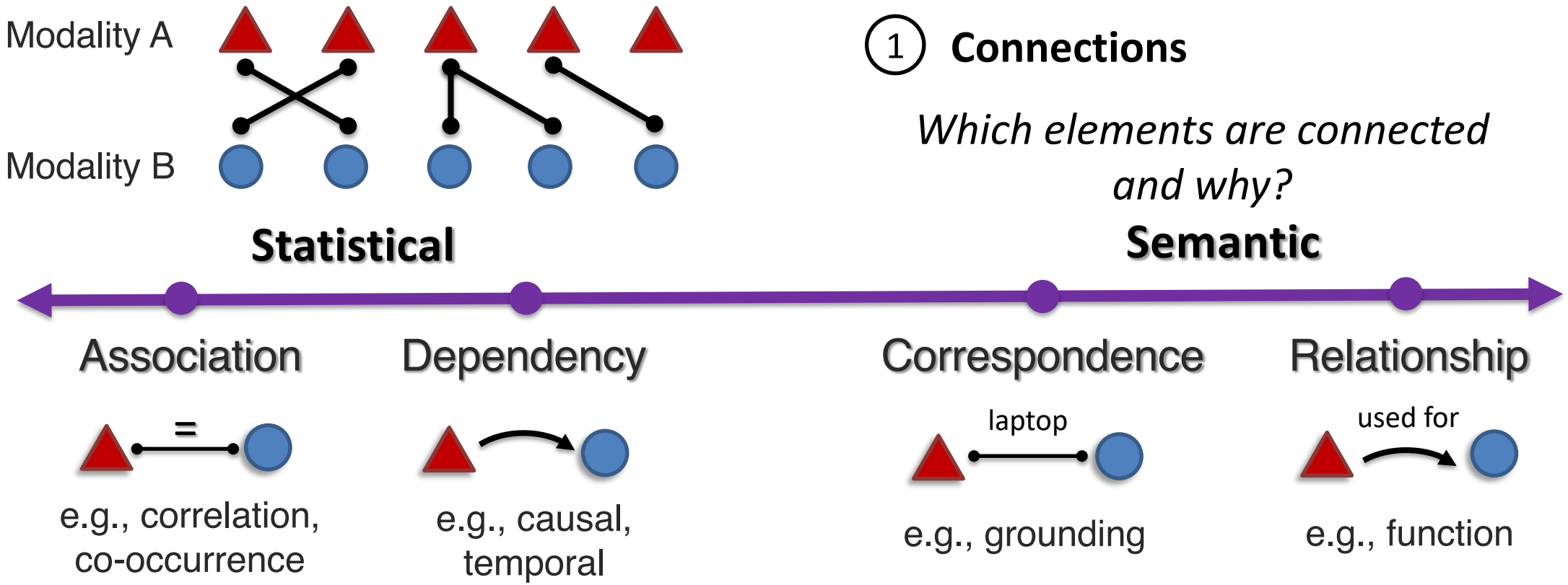


Translation model
Joint probabilistic model

[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 via Cyclic Translations Between Modalities. AAAI 2019]

Sub-Challenge 6b: Interconnections

Definition: Quantifying the presence and type of cross-modal connections and interactions in multimodal datasets and trained models.



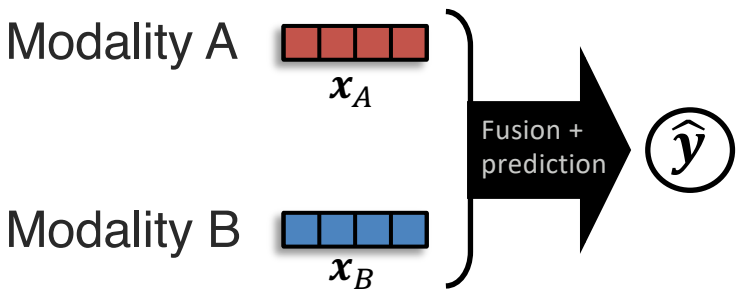
[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Quantifying Cross-modal Connections

Identifying overall presence of cross-modal connections

Statistical non-additive interactions [Friedman & Popescu, 2008, Sorokina et al., 2008]

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff f cannot be decomposed into a sum of unimodal subfunctions f_A, f_B such that $f(\boldsymbol{x}_A, \boldsymbol{x}_B) = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$.



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

If the additive projection $\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B)$ is equal to nonlinear fusion $f(\boldsymbol{x}_A, \boldsymbol{x}_B)$ then the non-additive interactions are not modeled.

μ measures **overall quantity** of cross-modal interactions on a trained model + dataset.

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

Quantifying Cross-modal Connections

Identifying individual cross-modal connections

Statistical non-additive interactions [Friedman & Popescu, 2008, Sorokina et al., 2008]

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff f cannot be decomposed into a sum of unimodal subfunctions f_A, f_B such that $f(\boldsymbol{x}_A, \boldsymbol{x}_B) = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$.

f exhibits interactions between 2 features $\boldsymbol{x}_A$ and $\boldsymbol{x}_B$ iff $\frac{\partial f^2}{\partial \boldsymbol{x}_A \partial \boldsymbol{x}_B} > 0$.

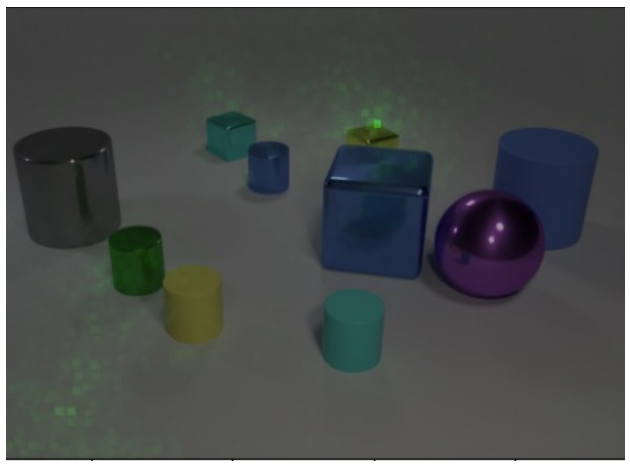
Natural second-order extension of gradient-based approaches!

[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Quantifying Cross-modal Connections

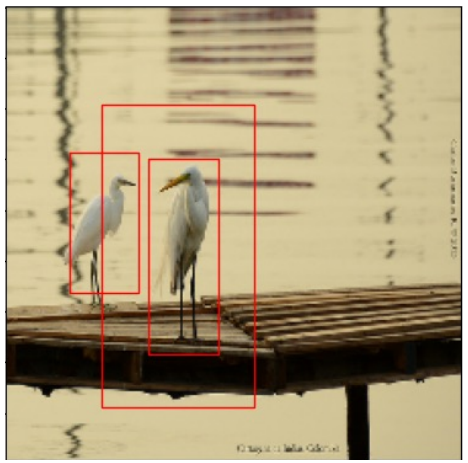
Identifying individual cross-modal connections

CLEVR



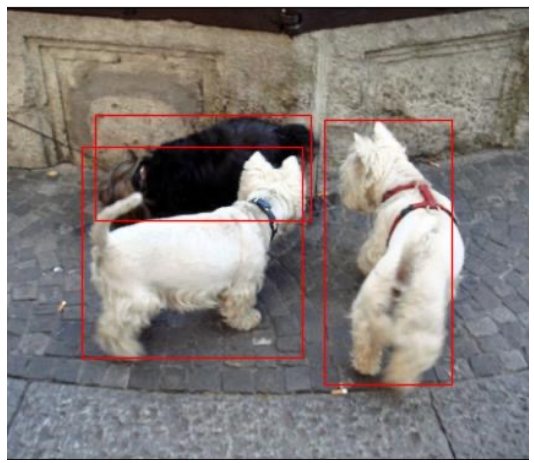
*The other small shiny thing that is the same shape as the **tiny yellow shiny object** is what color?*

VQA 2.0



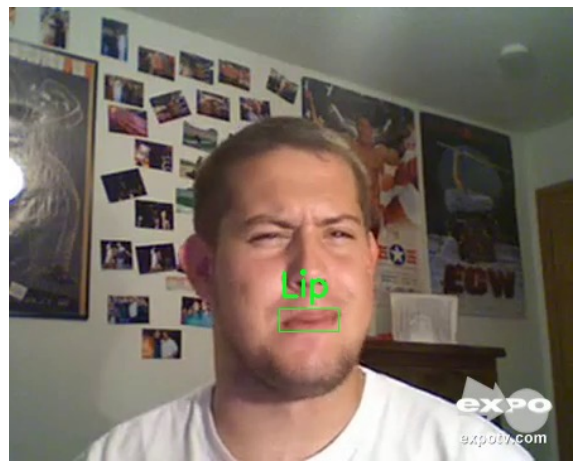
*How many **birds**?*

Flickr-30k



***Three small dogs**, two white and one black and white, on a sidewalk.*

CMU-MOSEI



*Why am I spending my money watching this? **(sigh)** I think I was more **sad**…*

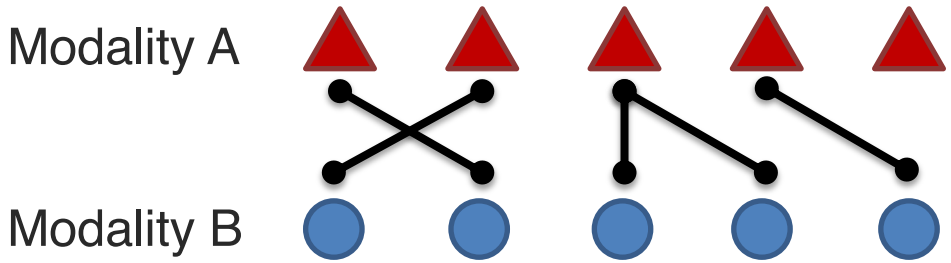
Correspondence

Relationship

[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

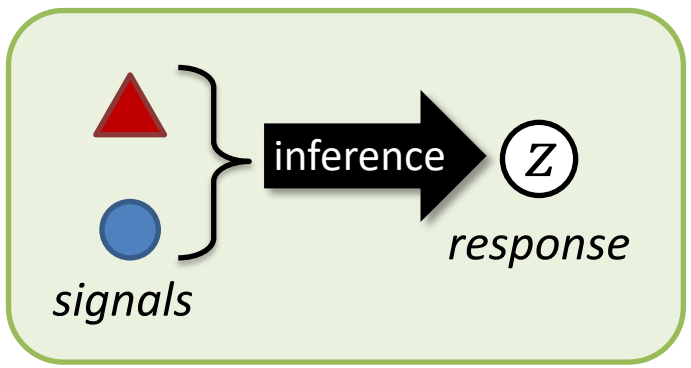
Sub-Challenge 6b: Interconnections

Definition: Quantifying the presence and type of cross-modal connections and interactions in multimodal datasets and trained models.



② **Cross-modal interactions**

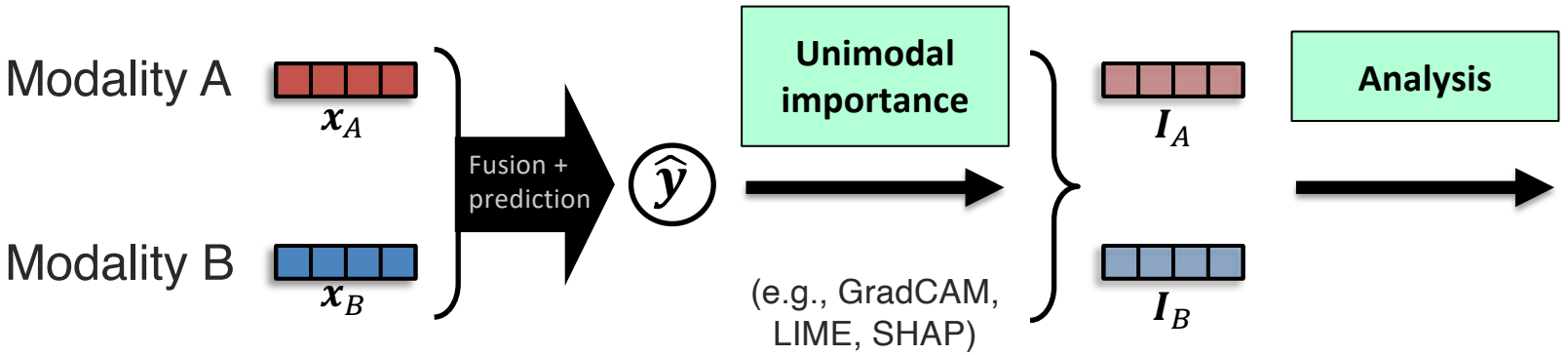
How are connected elements interacting during inference?

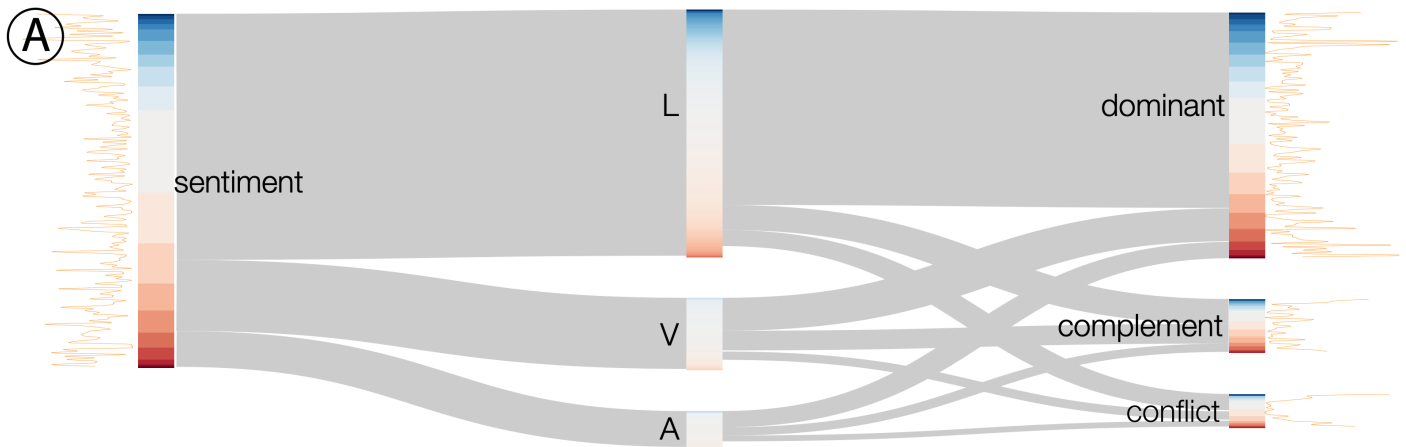


Interactions happen during inference!

Quantifying Cross-modal Interactions

Classification of cross-modal interactions





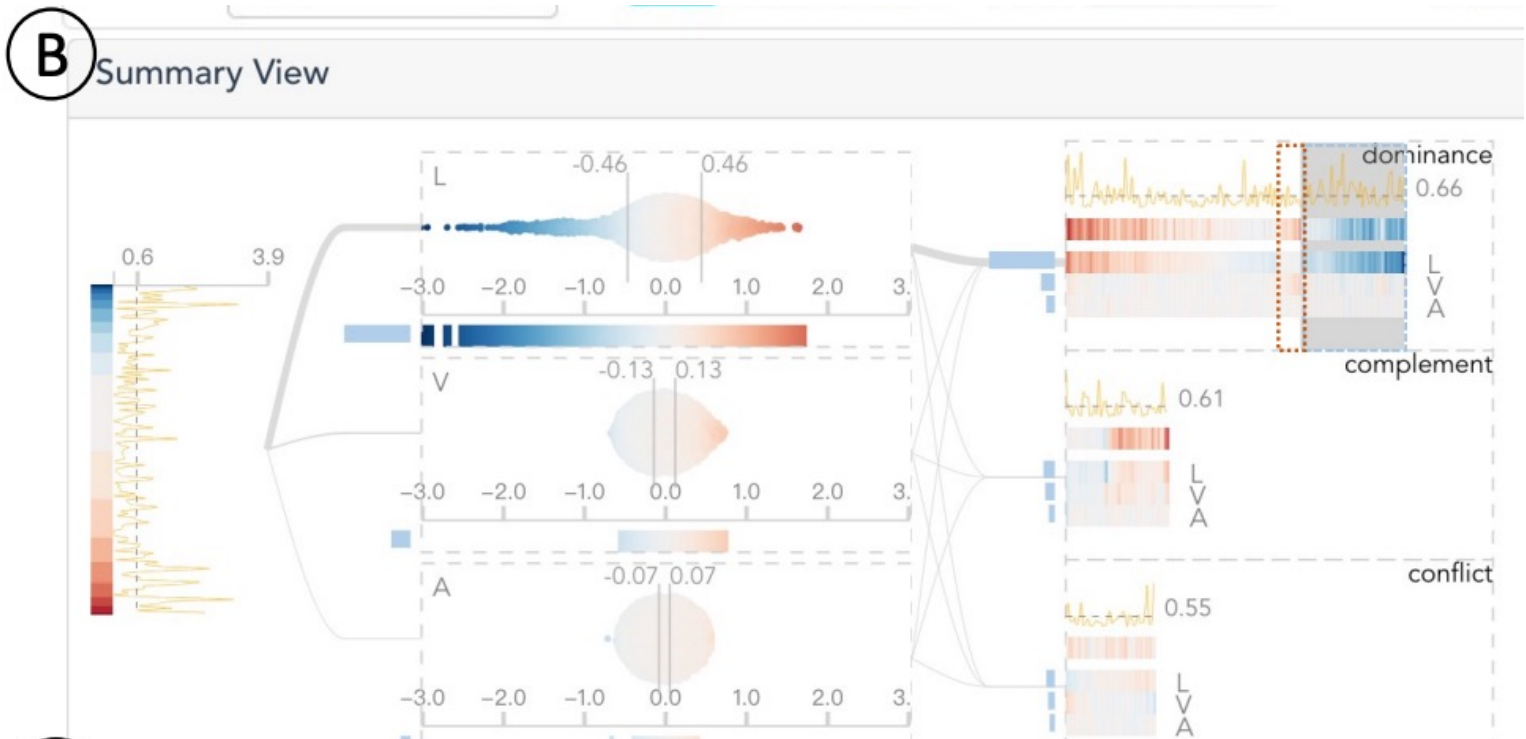
Language is often **dominant** in multimodal sentiment analysis

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Vis and Com Graphics 2021]

Quantifying Cross-modal Interactions

Visualization website

See interactive website: https://andy-xingbowang.com/m2lens/



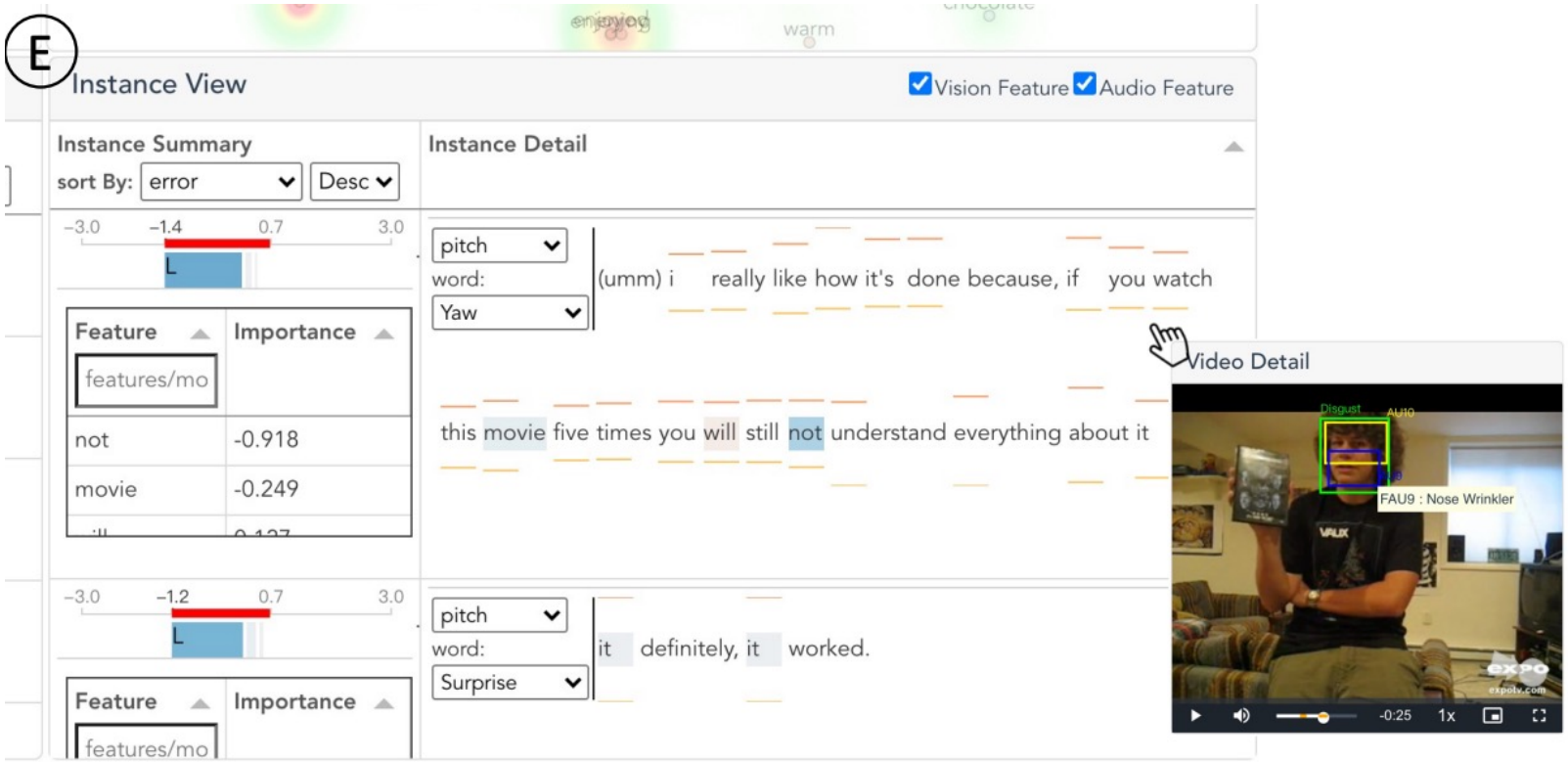
Summary of cross-modal interactions across entire dataset.

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Vis and Com Graphics 2021]

Quantifying Cross-modal Interactions

Visualization website

See interactive website: https://andy-xingbowang.com/m2lens/



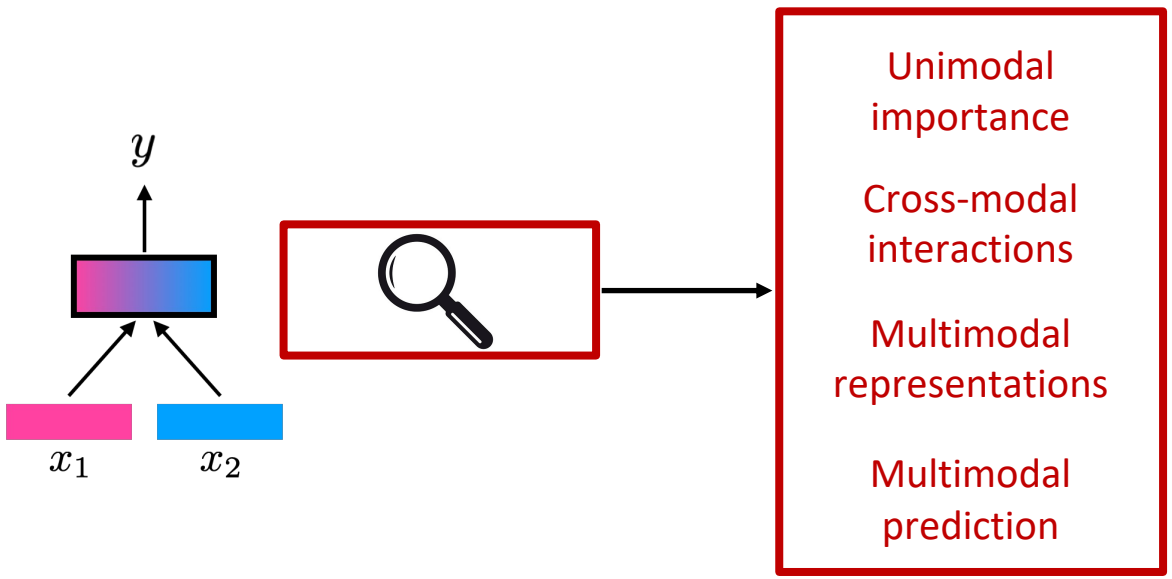
Summary of cross-modal interactions in a single instance.

[Wang et al., M2Lens: Visualizing and Explaining Multimodal Models for Sentiment Analysis. IEEE Trans Vis and Com Graphics 2021]

Evaluating Interpretability

How can we evaluate the success of interpreting internal mechanics?

Problem: real-world datasets and models do not have unimodal importance, cross-modal interactions, representations annotated!

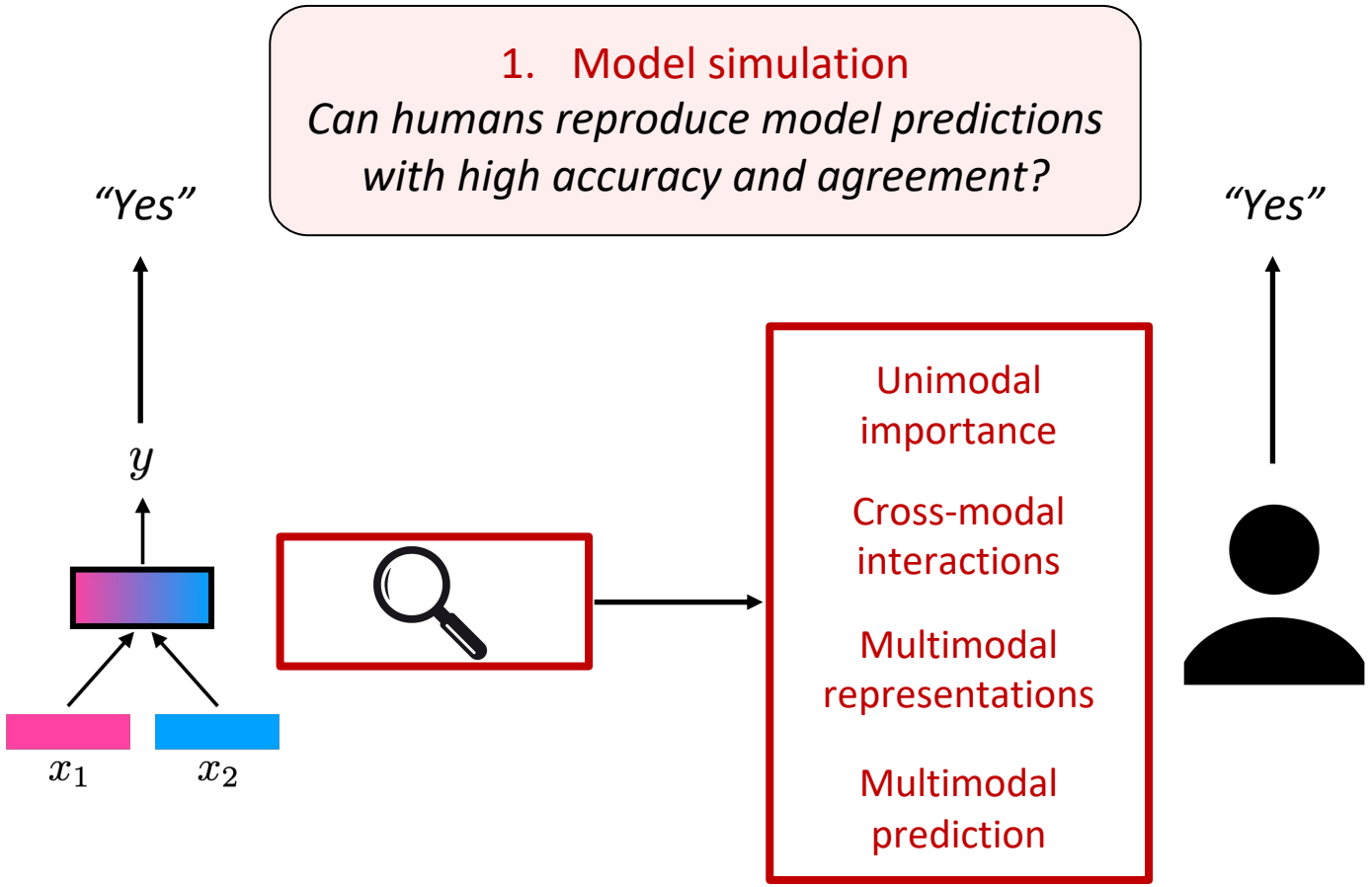


[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability

Open challenges

Model simulation



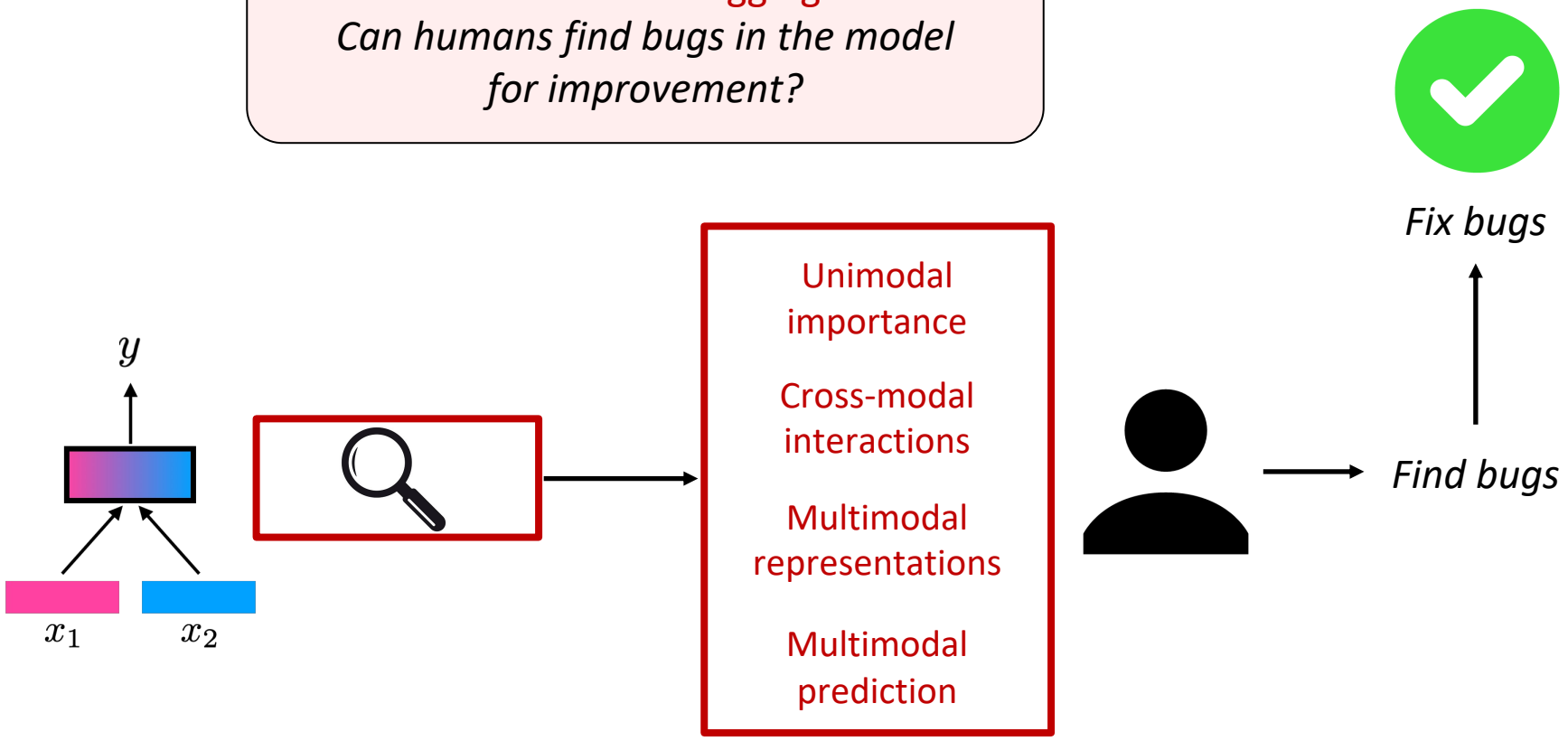
[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Evaluating Interpretability

Open challenges

Model error analysis and debugging

2. Model debugging
Can humans find bugs in the model for improvement?



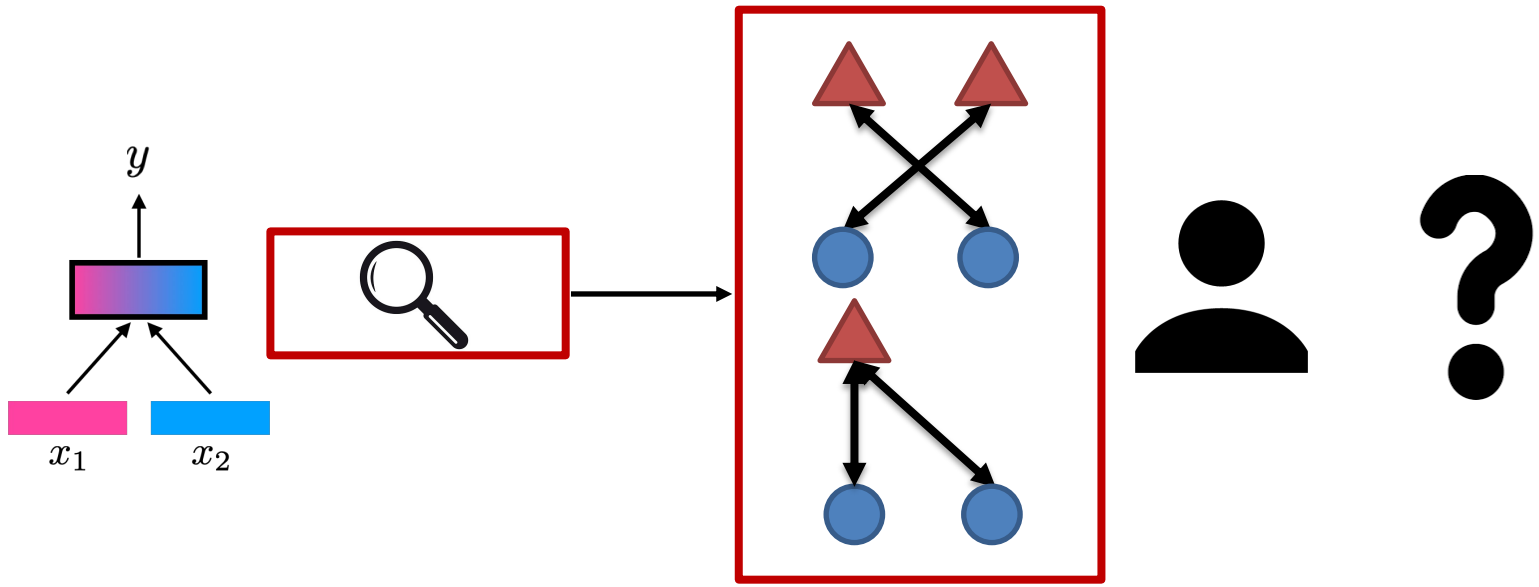
[Liang et al., MultiViz: An Analysis Benchmark for Visualizing and Understanding Multimodal Models. arXiv 2022]

Challenges

Open challenges

Open challenges:

- Faithfulness: do explanations accurately reflect model's internal mechanics?
- Usefulness: unclear if explanations help humans
- Disagreement: different interpretation methods may generate different explanations
- Evaluate: how to best evaluate interpretation methods

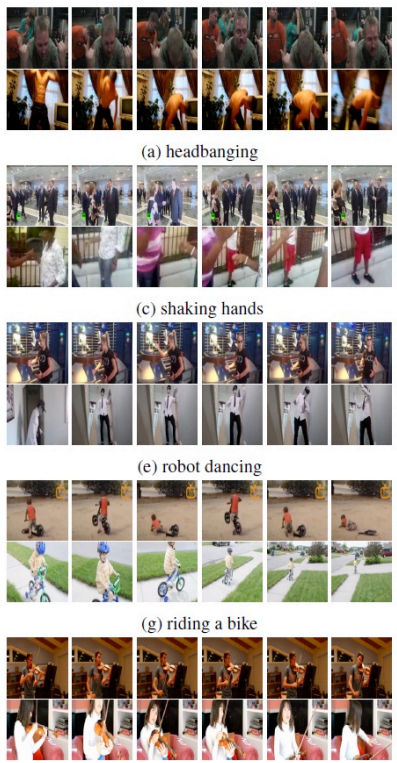


[Chandrasekaran et al., Do explanations make VQA models more predictable to a human? EMNLP 2018]
[Krishna et al., The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective. arXiv 2022]

Sub-Challenge 6c: Multimodal Learning Process

Definition: Characterizing the learning and optimization challenges involved when learning from heterogeneous data.

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]

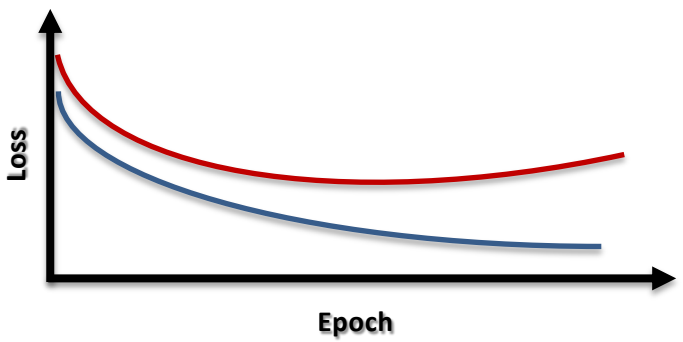
Optimization challenges

Open challenges

Learning and optimization challenges

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

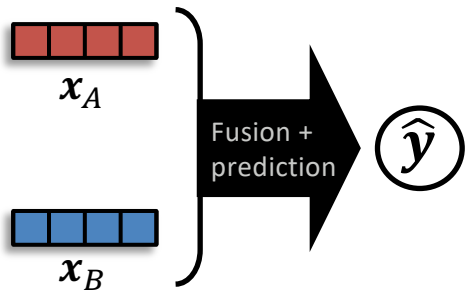
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

Optimization challenges

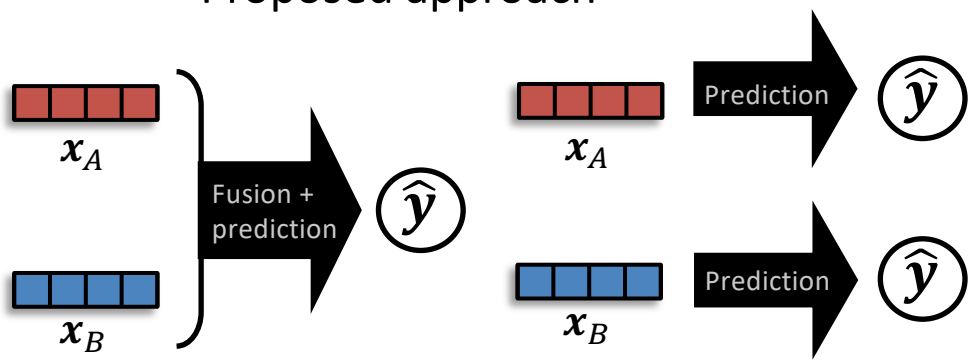
Open challenges

Learning and optimization challenges

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

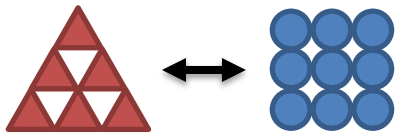
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]

More Quantification

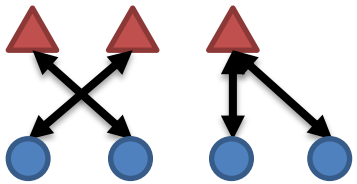
Dimensions of quantification

Representation **Alignment** **Reasoning** **Transference** **Generation**

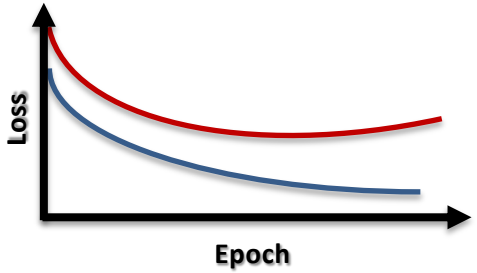
Heterogeneity



Interactions



Learning



Open challenges

Conclusion

What is Multimodal?

Why is it hard?

What is next?

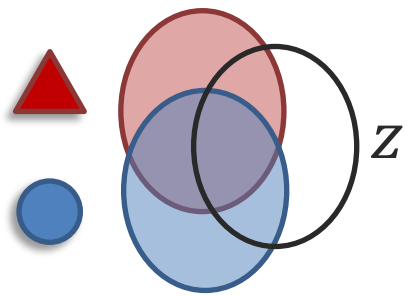
Heterogeneous

+

Connected

+

Interacting



Representation

Alignment

Reasoning

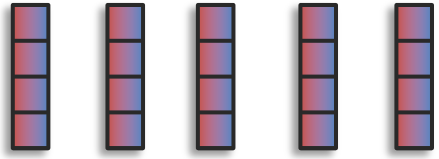
Generation

Transference

Quantification

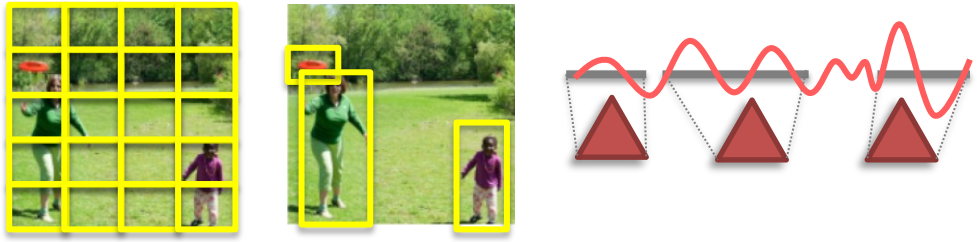
Future Direction: Heterogeneity

Homogeneity vs Heterogeneity



Examples:

Arbitrary Tokenization



Beyond Additive Interactions

Causal, logical interactions

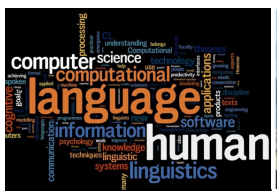
Brain-inspired representations

Future Direction: High-modality

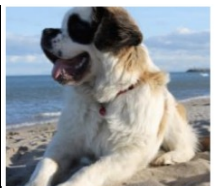
MultiBench

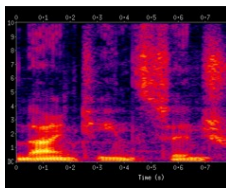
https://github.com/pliang279/MultiBench

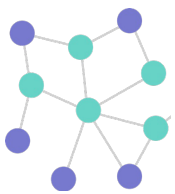
Few modalities → High-modality



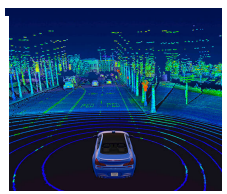
Language

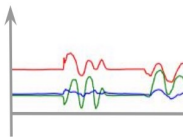

Vision


Audio


Graphs

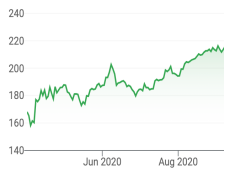

Control


LIDAR

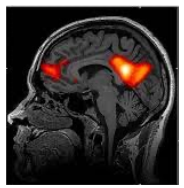

Sensors


Set

Table



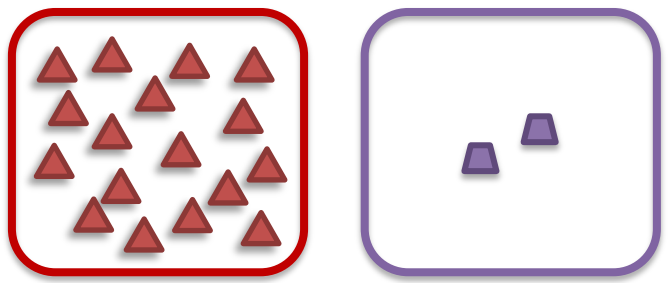
Financial


Medical

Examples:

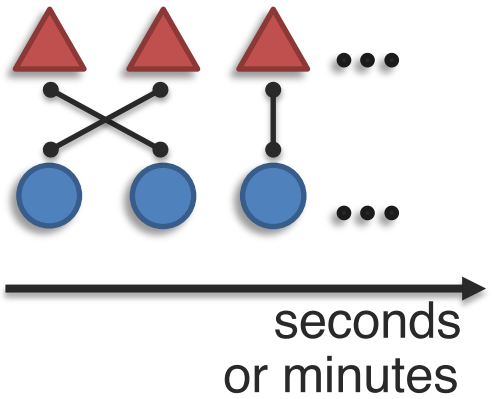
Non-parallel learning

Limited resources



Future Direction: Long-term

Short-term

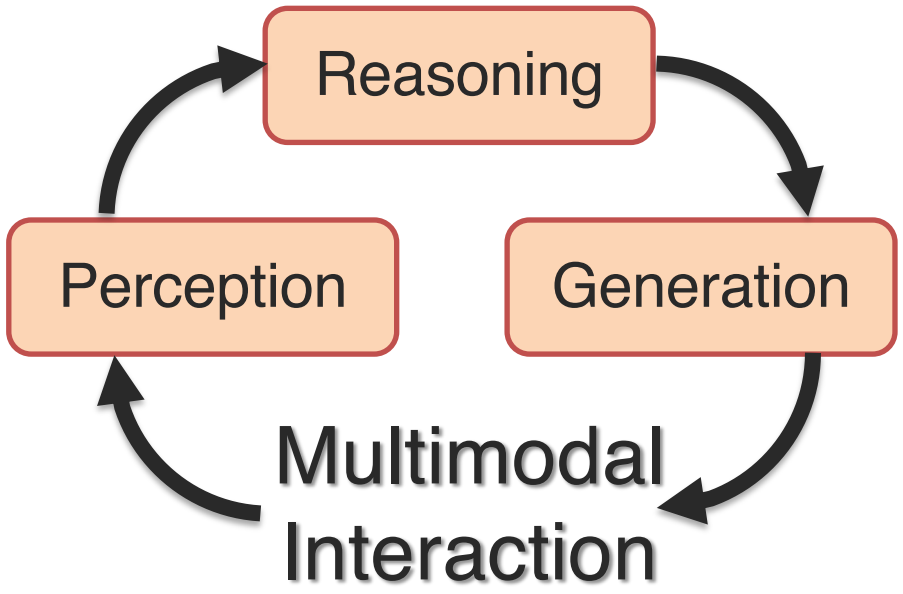


Long-term

Examples:

Compositionality **Memory** **Personalization**

Future Direction: Interaction



Social-IQ

https://www.thesocialiq.com/

Social Intelligence




Examples:

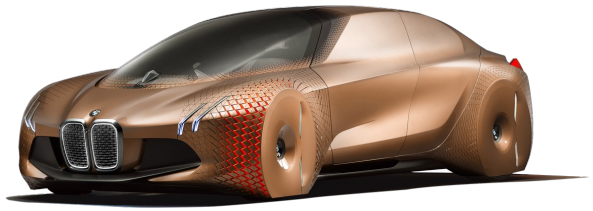
Multi-Party **Causality** **Ethical**

Future Direction: Real-world

MultiViz

https://github.com/pliang279/MultiViz

Healthcare Decision Support



Intelligent Interfaces and Vehicles

Online Learning and Education

Examples:

Robustness **Generalization** **Human-in-the-loop**

What is Multimodal?

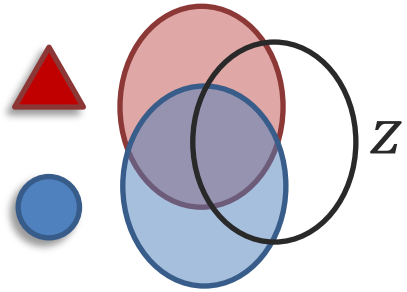
Heterogeneous

+

Connected

+

Interacting



Why is it hard?

Representation

Alignment

Reasoning

Generation

Transference

Quantification

What is next?

Heterogeneity

High-modality

Long-term

Interaction

Real-world

Multimodal Challenges – Surveys, Tutorials and Courses

2016

Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

(Arxiv 2017, IEEE TPAMI journal, February 2019)

https://arxiv.org/abs/1705.09406

Tutorials: CVPR 2016, ACL 2016, ICMI 2016, …

Graduate-level courses:

Multimodal Machine Learning (11th edition)
https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

Advanced Topics in Multimodal Machine Learning
https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

2022

Foundations and Recent Trends in Multimodal ML

Paul Liang, Amir Zadeh, and Louis-Philippe Morency

- 6 core challenges
- 50+ taxonomic classes
- 600+ referenced papers

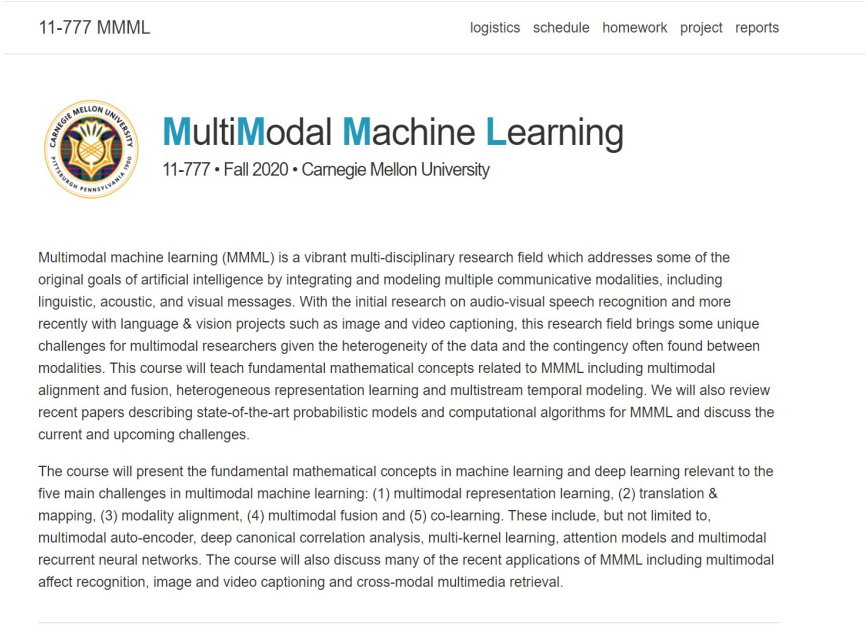
Tutorials: CVPR 2022, NAACL 2022, …

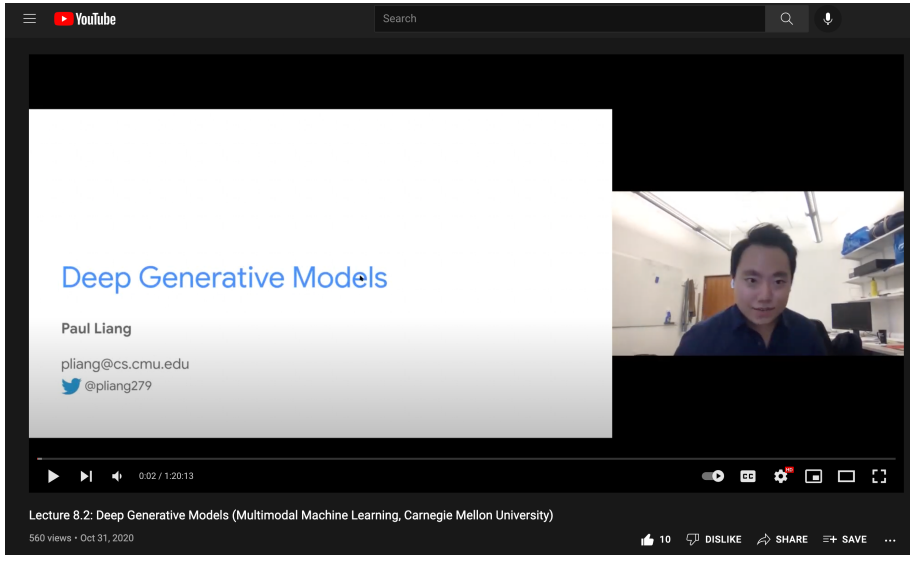
Updated graduate-level course:

Multimodal Machine Learning (12th edition)
Fall 2022 semester

[Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022]

Full Multimodal ML Course @ CMU





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

https://www.youtube.com/c/LPMorency/videos

Advanced Topics in Multimodal ML @ CMU

Advanced Topics in MultiModal Machine Learning

11-877 • Spring 2022 • Carnegie Mellon University

Multimodal machine learning (MMML) is a vibrant multi-disciplinary research field which addresses some of the original goals of artificial intelligence by integrating and modeling multiple communicative modalities, including language, vision, and acoustic. This research field brings some unique challenges for multimodal researchers given the heterogeneity of the data and the contingency often found between modalities. This course is designed to be a graduate-level course covering recent research papers in multimodal machine learning, including technical challenges with representation, alignment, reasoning, generation, co-learning and quantifications. The main goal of the course is to increase critical thinking skills, knowledge of recent technical achievements, and understanding of future research directions.

- **Time:** Friday 10:10-11:30 am
- **Location:** Virtual for the first 2 weeks (find zoom link in piazza), GHC 5222 thereafter
- **Discussion and Q&A:** Piazza
- **Assignment submissions:** Canvas (for registered students only)
- **Contact:** Students should ask all course-related questions on Piazza, where you will also find announcements.

Instructor Louis-Philippe Morency
Email: morency@cs.cmu.edu

Instructor Amir Zadeh
Email: abagherz@cs.cmu.edu

Instructor Paul Liang
Email: pliang@cs.cmu.edu

1/28 Week 2: **Cross-modal interactions** [synopsis]

- What are the different ways in which modalities can interact with each other in multimodal tasks? Can we formalize a taxonomy of such cross-modal interactions, which will enable us to compare and contrast them more precisely?
- What are the design decisions (aka inductive biases) that can be used when modeling these cross-modal interactions in machine learning models?
- What are the advantages and drawbacks of designing models to capture each type of cross-modal interaction? Consider not just prediction performance, but tradeoffs in time/space complexity, interpretability, etc.
- Given an arbitrary dataset and prediction task, how can we systematically decide what type of cross-modal interactions exist, and how can that inform our modeling decisions?
- Given trained multimodal models, how can we understand or visualize the nature of cross-modal interactions?

- Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!
- What Does BERT with Vision Look At?
- Multiplicative Interactions and Where to Find Them
- Cooperative Learning for Multi-view Analysis
- Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers
- Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks

2/4 Week 3: **Multimodal co-learning** [synopsis]

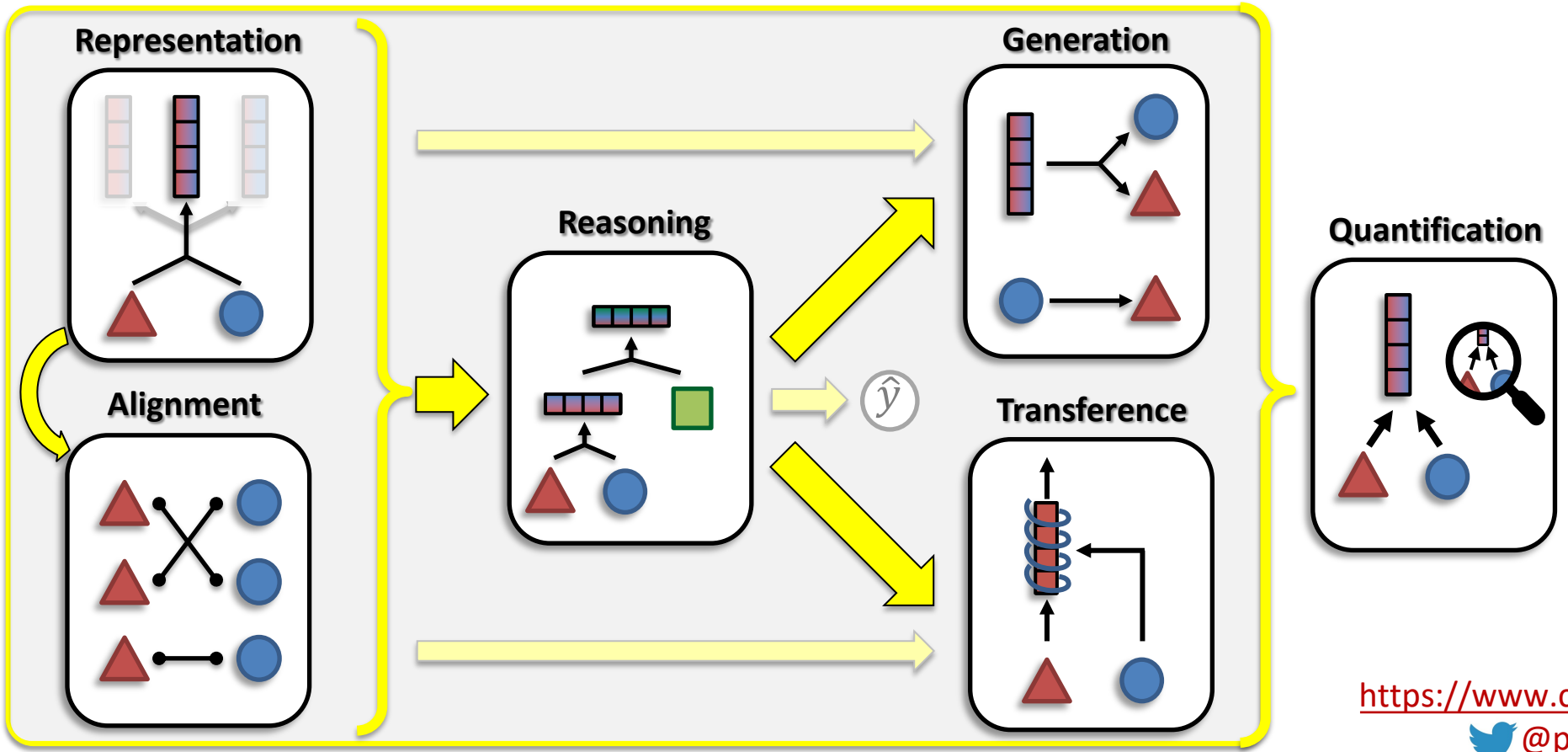
- What are the types of cross-modal interactions involved to enable such co-learning scenarios where multimodal training ends up generalizing to unimodal testing?
- What are some design decisions (inductive bias) that could be made to promote transfer of information from one modality to another?
- How do we ensure that during co-learning, only useful information is transferred, and not some undesirable bias? This may become a bigger issue in low-resource settings.
- How can we know if co-learning has succeeded? Or failed? What approaches could we develop to visualize and probe the success of co-learning?
- How can we formally, empirically, or intuitively measure the additional information provided by auxiliary modality? How can we design controlled experiments to test these hypotheses?
- What are the advantages and drawbacks of information transfer during co-learning? Consider not just prediction performance, but also tradeoffs with complexity, interpretability, fairness, etc.

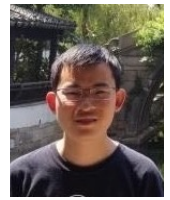
- Multimodal Prototypical Networks for Few-shot Learning
- SMIL: Multimodal Learning with Severely Missing Modality
- Multimodal Co-learning: Challenges, Applications with Datasets, Recent Advances and Future Directions
- Vokenization: Improving Language Understanding with Contextualized, Visual-Grounded Supervision
- What Makes Multi-modal Learning Better than Single (Provably)
- Found in Translation: Learning Robust Joint Representations by Cyclic Translations Between Modalities
- Zero-Shot Learning Through Cross-Modal Transfer
- 12-in-1: Multi-Task Vision and Language Representation Learning
- A Survey of Reinforcement Learning Informed by Natural Language

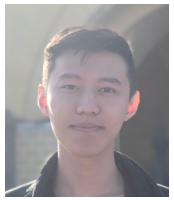
https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

THE END – Questions?

Multimodal is the science of
heterogeneous and **interconnected** data ☺




Yiwei Lyu


Peter Wu


Catherine Cheng


Amir Zadeh


LP Morency


Ruslan Salakhutdinov


Manuel Blum


Lenore Blum

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

@pliang279