Multimodal Machine Learning

Lecture 3.2: Multimodal Representations (Part 1)

Louis-Philippe Morency

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

Administrative Stuff

Reading Assignments – Reminder

Week 3 reading assignment was posted

1. **Friday 8pm:** Post your summary
2. **Monday 8pm:** End of the reading assignment

Be sure to post your discussion comments before Monday 8pm!

➡ Start the discussion early ☺

➡ Late submissions will be accounted

Primary TAs

- Each team will have one primary TA
- Meetings with primary TA will be scheduled for next week
  - Feedback for the pre-proposals
- Contact your primary TA anytime (piazza or email)
  - Groups will be created in Piazza for each team
- Some projects may have a secondary TA, with complementary expertise

First Project Assignment

Due date: Sunday 9/25 at 8m

Four main sections:

- Introduction
- Related work
- Experimental setup
- Research ideas

Follows ICML paper format

The two main sections are related work and research ideas

teammates = # research ideas

Page limit depends on team size:

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

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 3.2: Multimodal Representations (Part 1)

Louis-Philippe Morency

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

Lecture Objectives

- Multimodal representations
  - Cross-modal interactions
- Representation fusion
  - Additive and multiplicative fusion
    - Tensor and polynomial fusion
  - Gated fusion
    - Modality-shift fusion
    - Dynamic fusion
  - Fusion on raw modalities
  - Multimodal autoencoder
- Measuring non-additive interactions

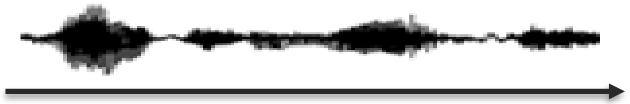
Multimodal Representation

Multimodal Machine Learning

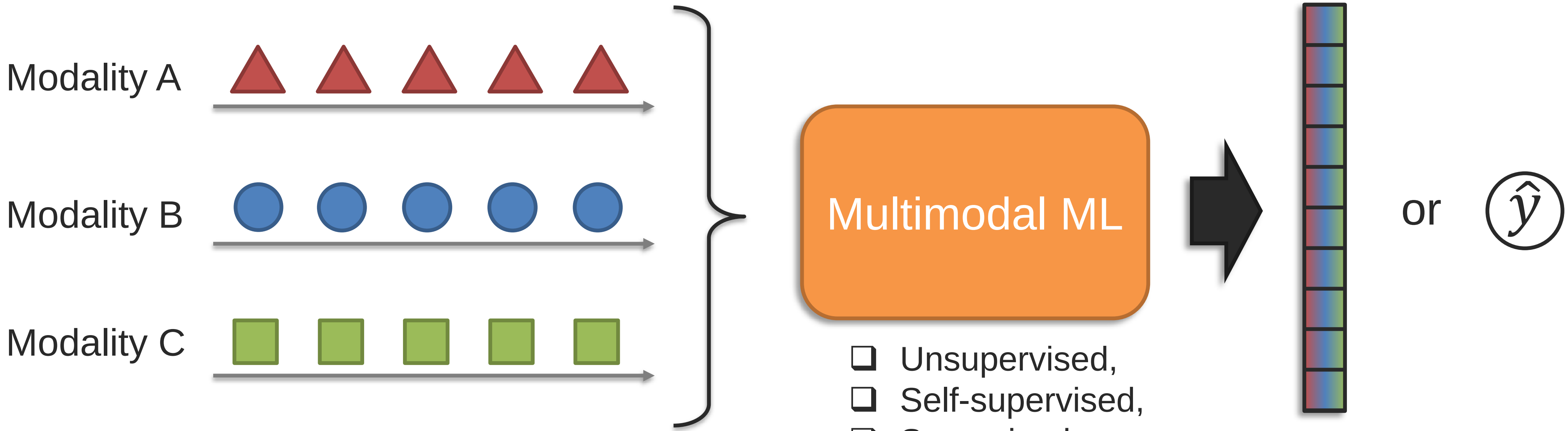
Language I really like this tutorial

Vision

Acoustic



Multimodal Machine Learning



- ❑ Unsupervised,
- ❑ Self-supervised,
- ❑ Supervised,
- ❑ Reinforcement,

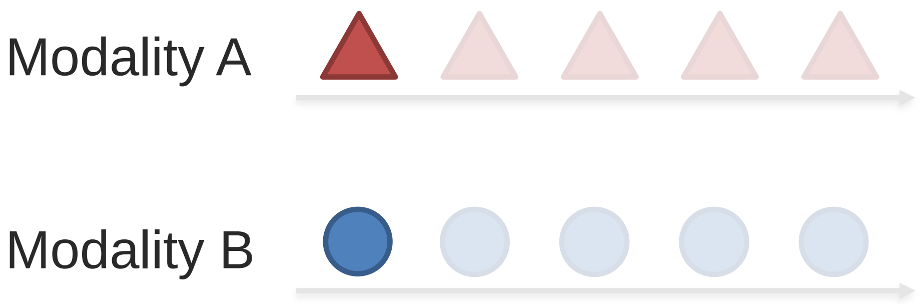
…

Challenge 1: Representation

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

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

Individual elements:



It can be seen as a "local" representation

or

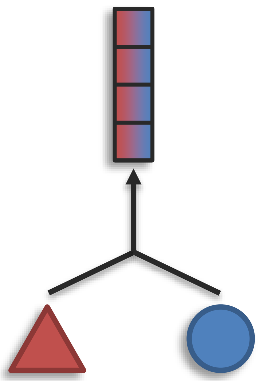
representation using holistic features

Challenge 1: Representation

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

Sub-challenges:

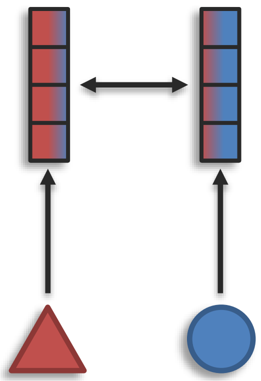
Fusion



modalities > # representations

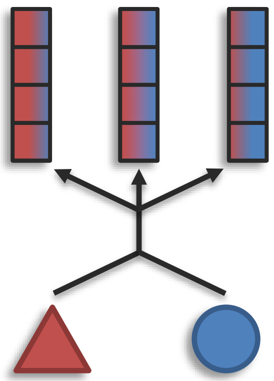
Today

Coordination



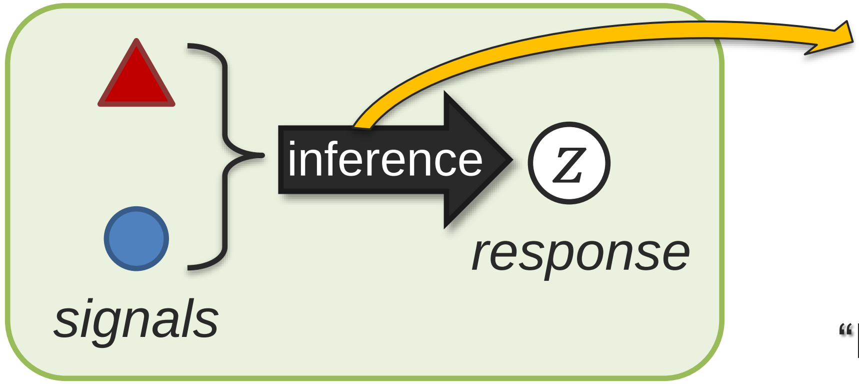
modalities = # representations

Fission



modalities < # representations

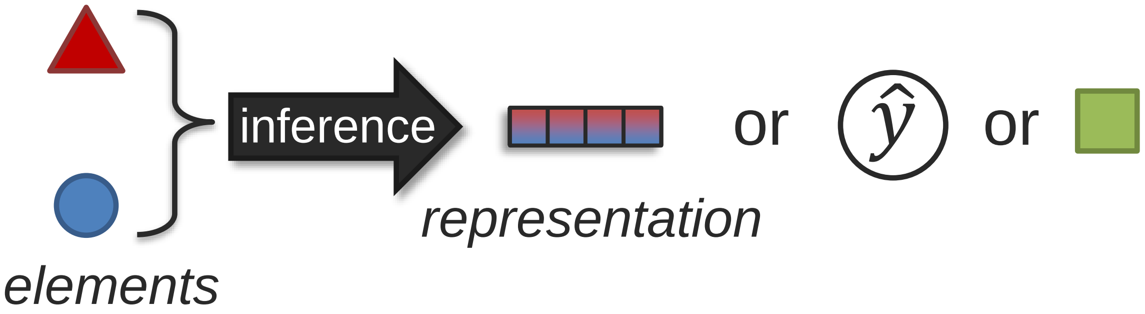
Cross-modal Interactions



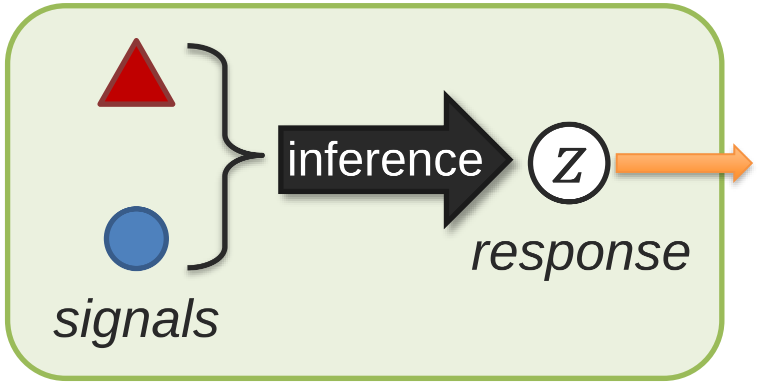
Interactions happen during inference!

"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation

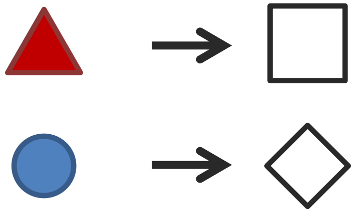


Interconnected Modalities



Types of interaction responses?
(a taxonomy)

Unimodal
Non-redundancy



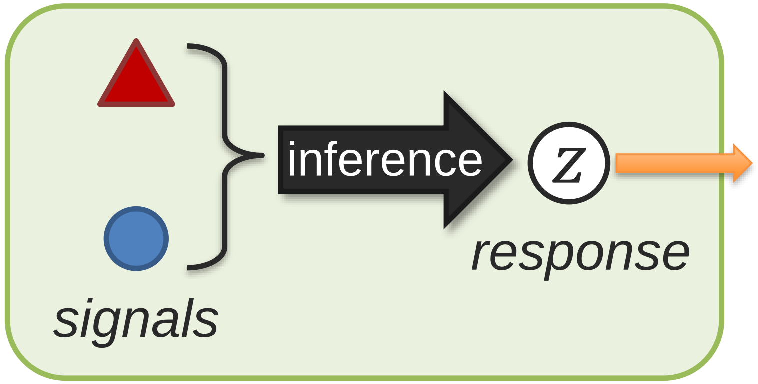
Is this a living room?

inference → *Yes!*

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

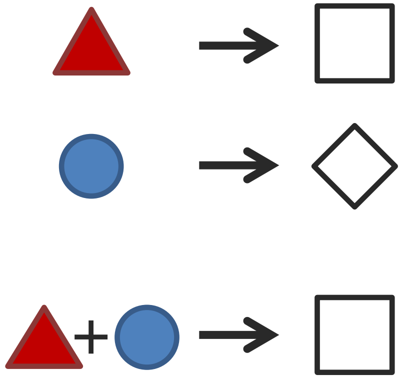
inference → *No, probably study room.*

Interconnected Modalities



Types of interaction responses?
(a taxonomy)

Unimodal
Non-redundancy



Multimodal
dominance

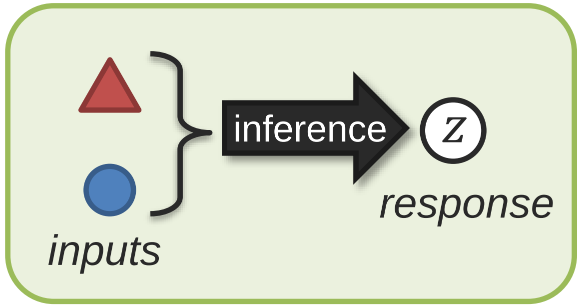
Is this a living room?

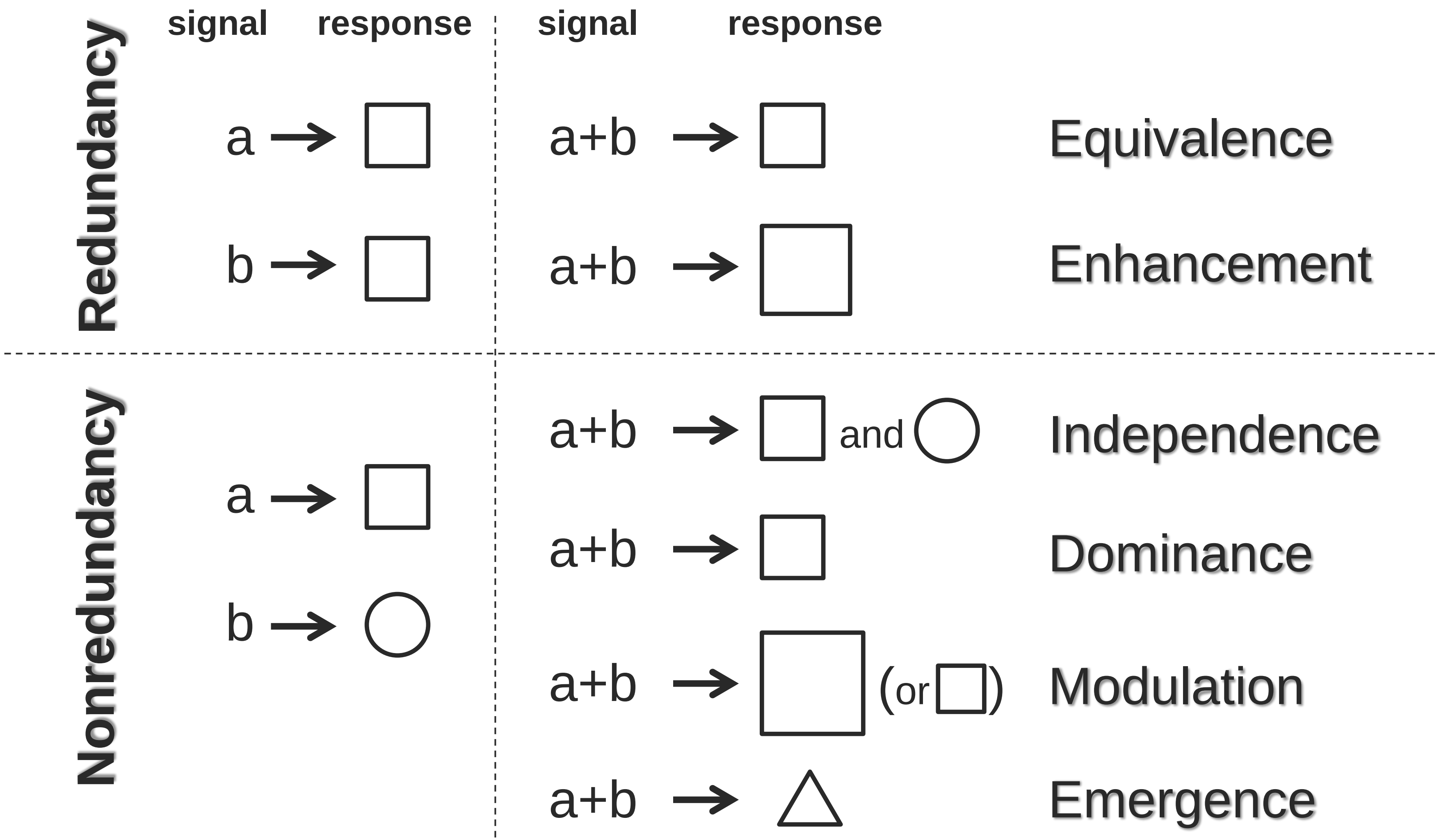
A teacup on the right of a laptop in a clean room.

inference

Yes!

Taxonomy of Interaction Responses – A Behavioral Science View

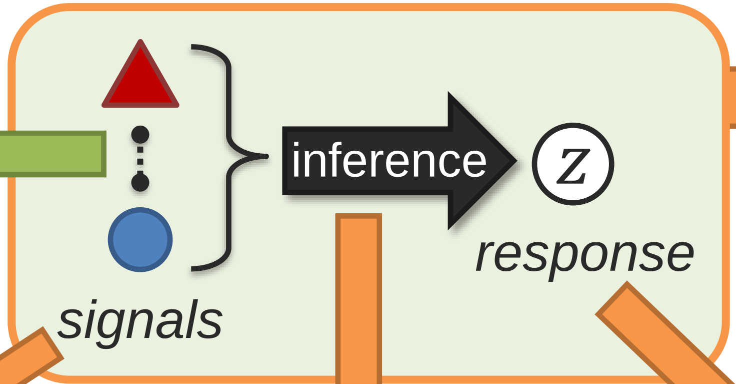




Multimodal Communication

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

Cross-modal Interactions – A Taxonomy



5 Connections

- Association
- Dependency
- Correspondence
- Relationship

...

4 Context

- Structure context
- Task relevance
- Context dependence

...

2 Modalities

- Unimodal
- Bimodal
- Trimodal
- High-modal,

...

3 Interactions

- Additive
- Multiplicative
- Polynomial
- Gated
- Nonlinear

…

1 Responses

- Redundancy
- Non-redundancy
- Dominance
- Emergence

…

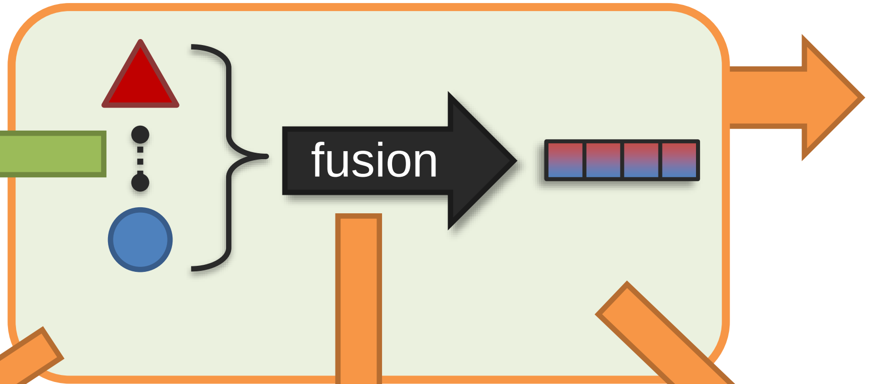
Cross-modal Interactions – Representation Fusion

Next week

5 Connections

- Association
- Dependency
- Correspondence
- Relationship

...



4 Context

- Structure context
- Task relevance
- Context dependence

...

2 Modalities

- Unimodal
- Bimodal
- Trimodal
- High-modal,

...

3 Interactions

- Additive
- Multiplicative
- Polynomial
- Gated
- Nonlinear

…

Today

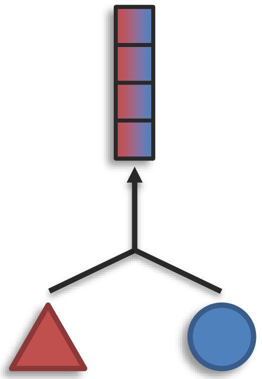
1 Responses

- Redundancy
- Non-redundancy
- Dominance
- Emergence

…

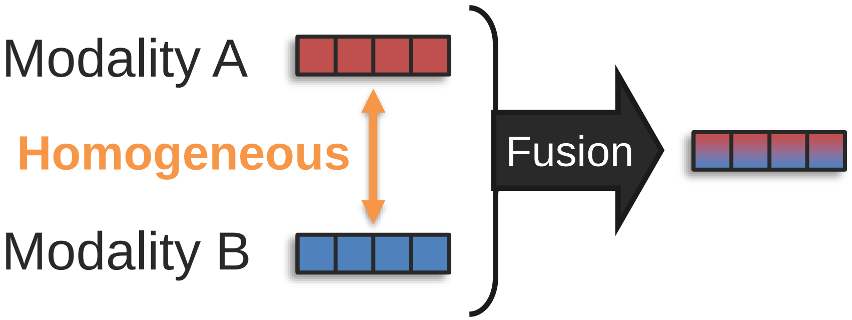
Representation Fusion

Sub-Challenge 1a: Representation Fusion

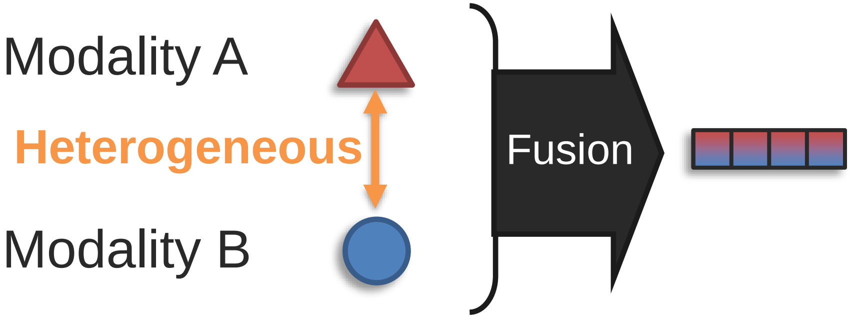


Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

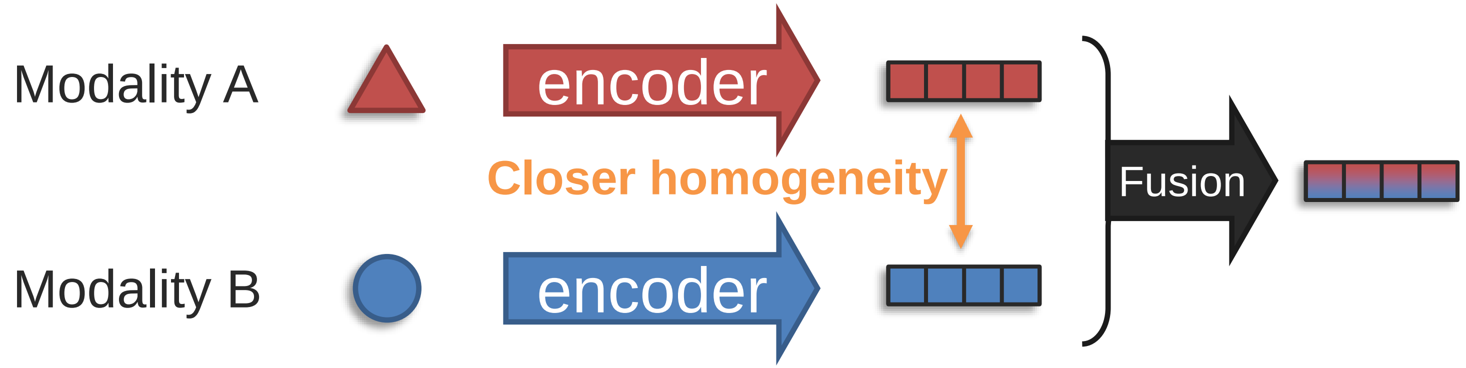
Basic fusion:



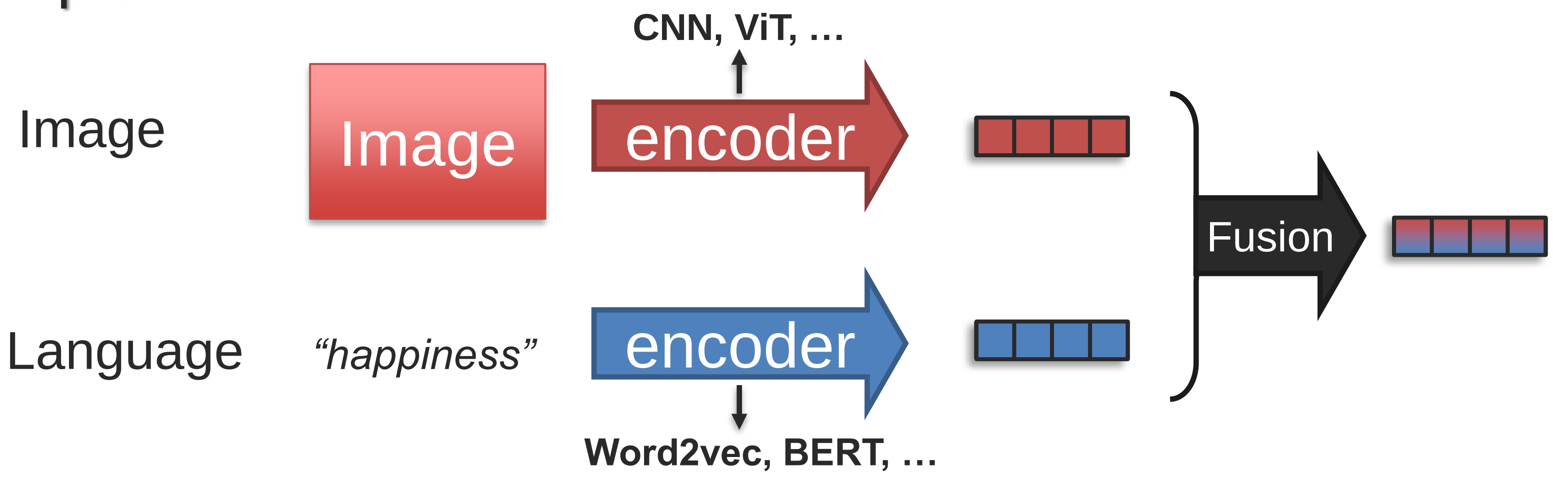
Raw-modality fusion:



Fusion with Unimodal Encoders



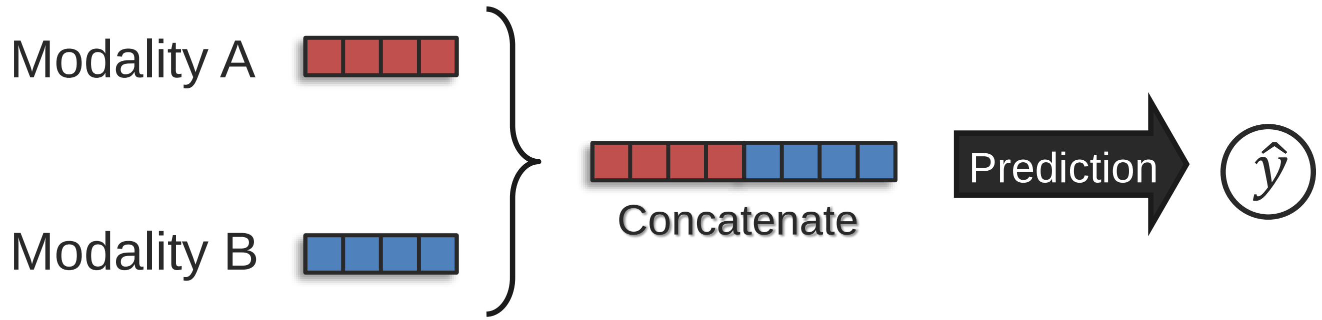
Example:



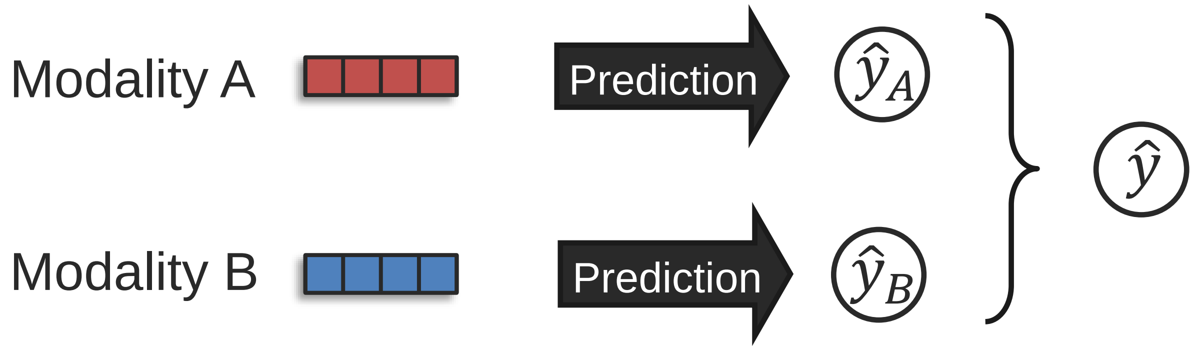
➡ Unimodal encoders can be jointly learned with fusion network, or pre-trained

Early and Late Fusion – A historical View

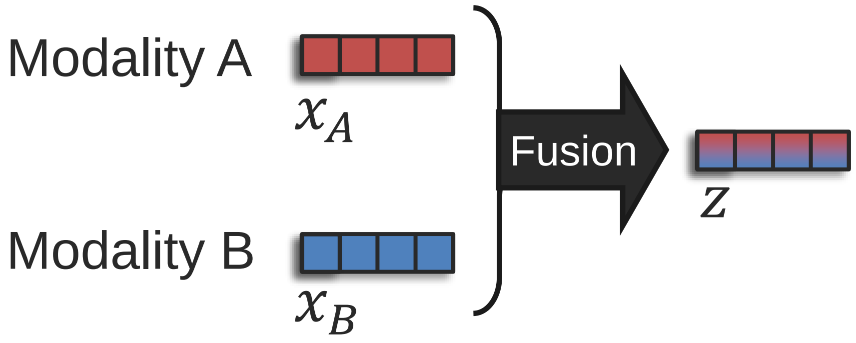
Early fusion:



Late fusion:



Basic Concepts for Representation Fusion (aka, Basic Fusion)



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

➡ Let's study the univariate case first
(only 1-dimensional features)

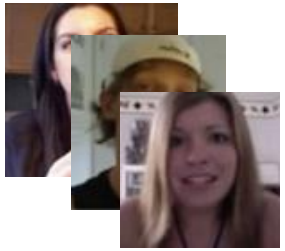
Linear regression:

$$z = \underbrace{w_0}_{\substack{\text{intercept}\\\text{(bias term)}}} + \underbrace{w_1 x_A + w_2 x_B}_{\substack{\text{Additive}\\\text{terms}}} + \underbrace{w_3 (x_A \times x_b)}_{\substack{\text{Multiplicative}\\\text{term}}} + \underbrace{\epsilon}_{\substack{\text{error}\\\text{(residual term)}}}$$

Linear Regression

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

300 book reviews



y: audience score

x_A: percentage of smiling

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

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

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

Linear regression:

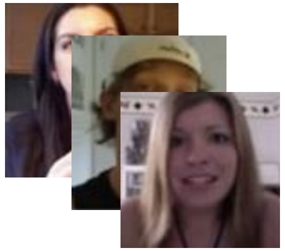
$$y = \underbrace{w_0}_{\text{intercept (bias term)}} + \underbrace{w_1 x_A + w_2 x_B}_{\text{Additive terms}} + \underbrace{w_3 (x_A \times x_b)}_{\text{Multiplicative term}} + \underbrace{\epsilon}_{\text{error (residual term)}}$$

- w_0: average score when x_A and x_B are zero
- w_1: effect from x_A variable only
- w_2: effect from x_B variable only
- w_3: effect from x_A and x_B interaction only
- ϵ: residual not modeled by w_0, w_1, w_2 or w_3

Linear Regression

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

300 book reviews



y: audience score

x_A: percentage of smiling

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

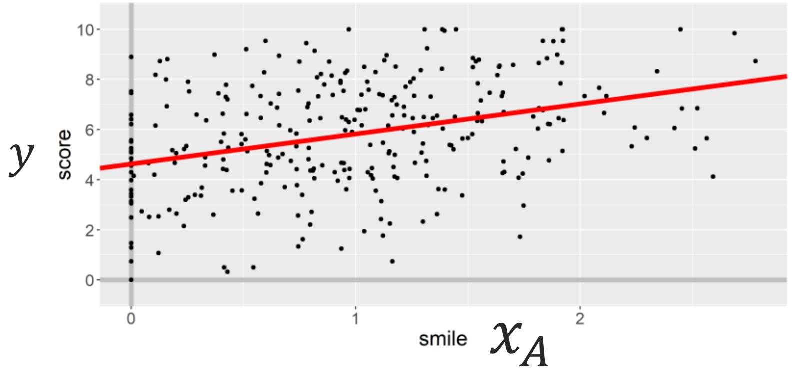
H1: *Does smiling reveal what the audience score was?*

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

Linear regression:

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

slope



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

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

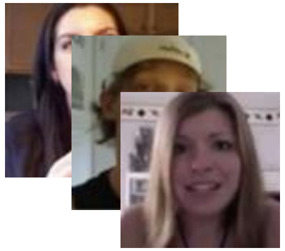
Confidence interval does not contain 0, so effect is significant

p-values would be another way to test hypothesis

Linear Regression

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

300 book reviews



y: audience score

x_A: percentage of smiling

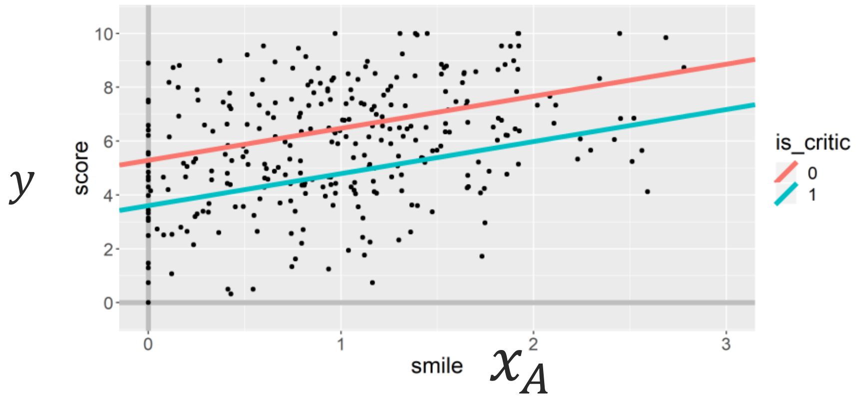
x_B: professional status
(0=non-critic, 1=critic)

H1: Does smiling reveal what the audience score was?

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

Linear regression:

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



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

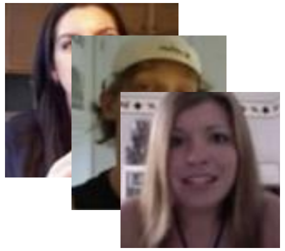
Positive effect

Negative effect

Linear Regression

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

300 book reviews



y: audience score

x_A: percentage of smiling

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

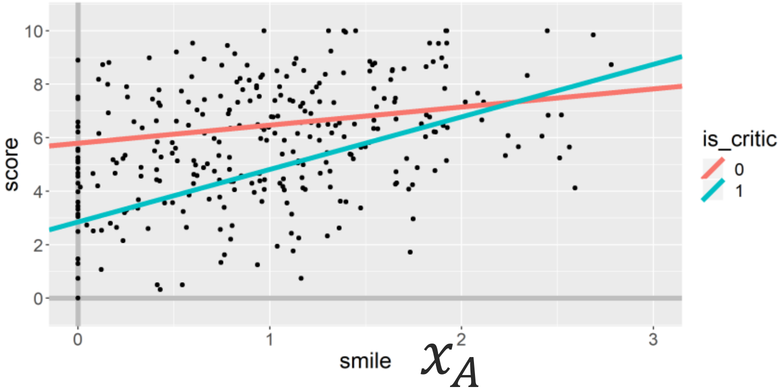
H1: *Does smiling reveal what the audience score was?*

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

Linear regression:

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

y

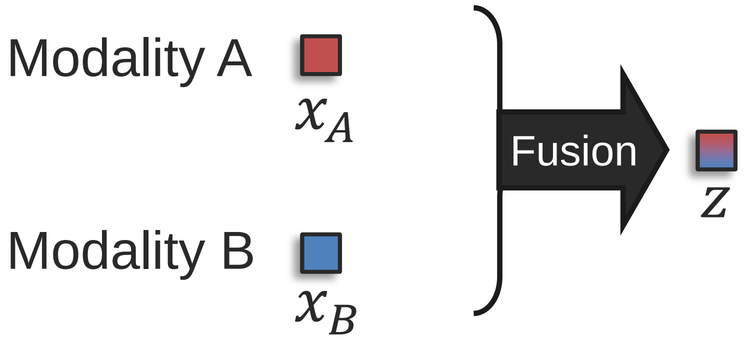


x_A

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

Multiplicative interaction!

Basic Concepts for Representation Fusion (aka, Basic Fusion)



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

Let's study the univariate case first

(only 1-dimensional features)

Linear regression:

$$z = \underbrace{w_0}_{\substack{\text{intercept} \\ \text{(bias term)}}} + \underbrace{w_1 x_A + w_2 x_B}_{\substack{\text{Additive} \\ \text{terms}}} + \underbrace{w_3 (x_A \times x_b)}_{\substack{\text{Multiplicative} \\ \text{term}}} + \underbrace{\epsilon}_{\substack{\text{error} \\ \text{(residual term)}}}$$

1. Additive terms:

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

2. Multiplicative "interaction" term:

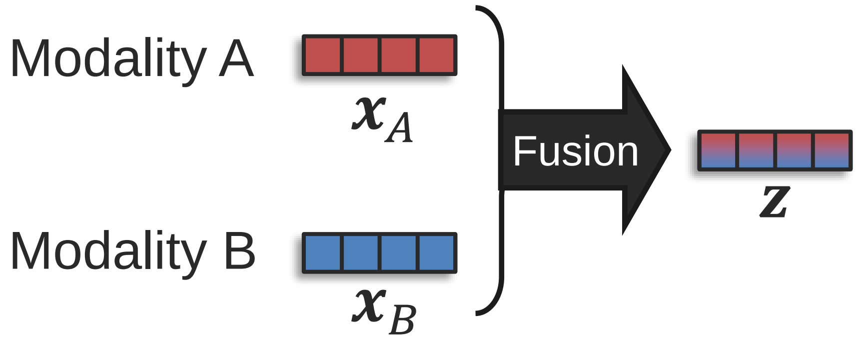
$$z = w_3 (x_A \times x_b) + \epsilon$$

3. Additive and multiplicative terms:

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

Additive Fusion ➡ Back to multivariate case!

↳(multi-dimensional features)

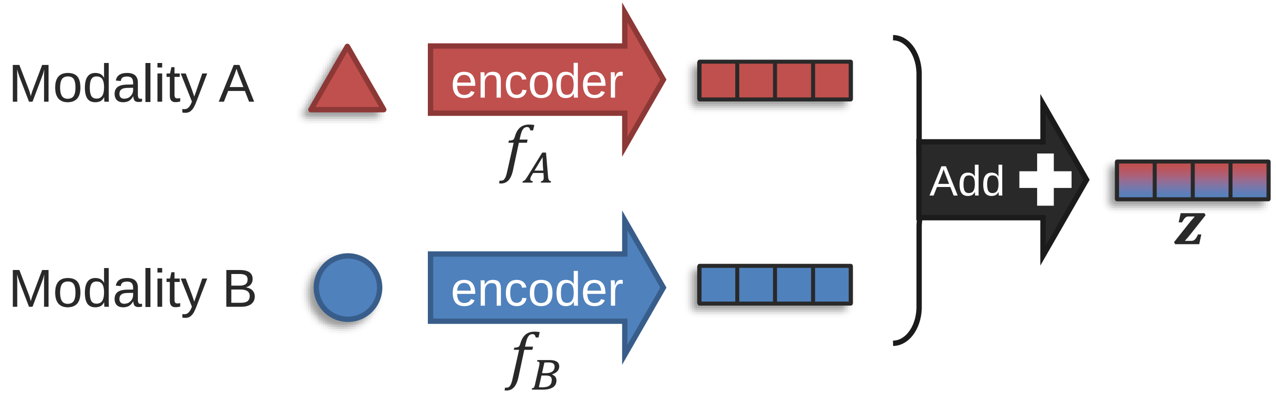


Additive fusion:

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

➡ 1-layer neural network can be seen as additive

With unimodal encoders:



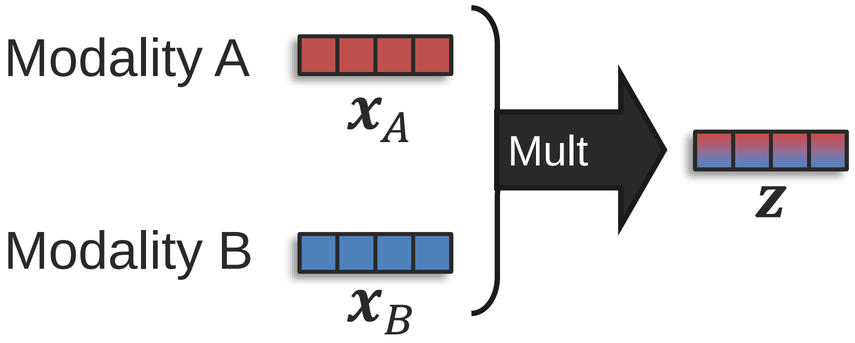
Additive fusion:

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

➡ It could be seen as an ensemble approach

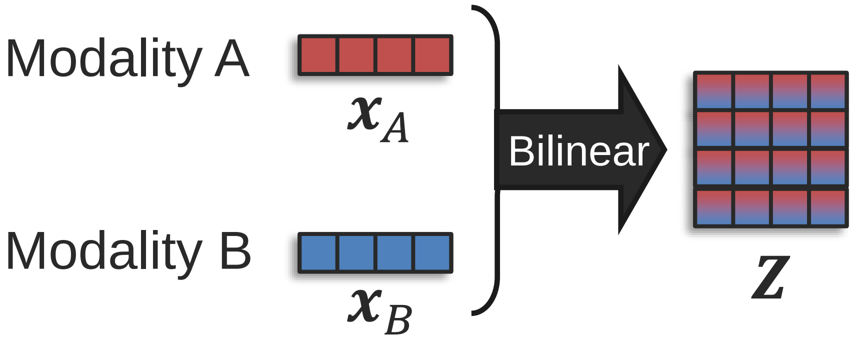
(late fusion)

Multiplicative Fusion



Simple multiplicative fusion:

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

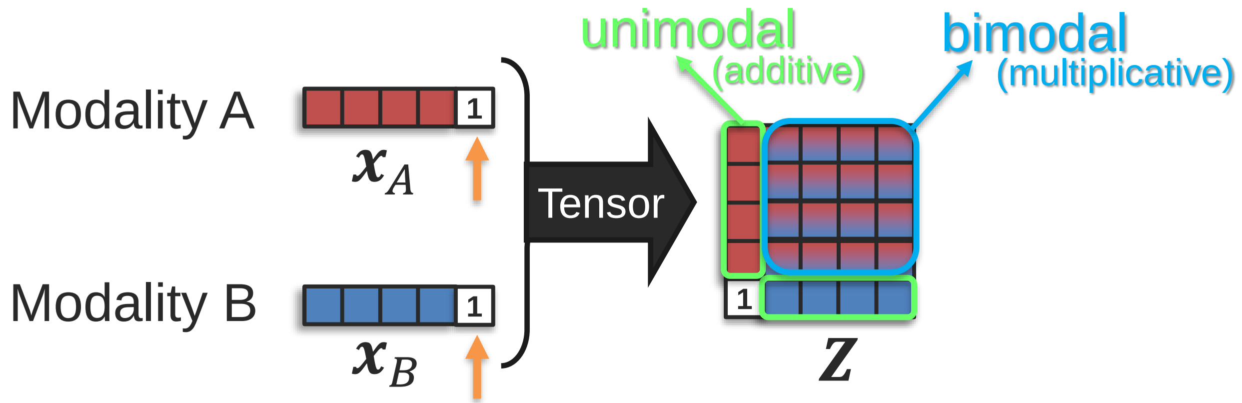


Bilinear Fusion:

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

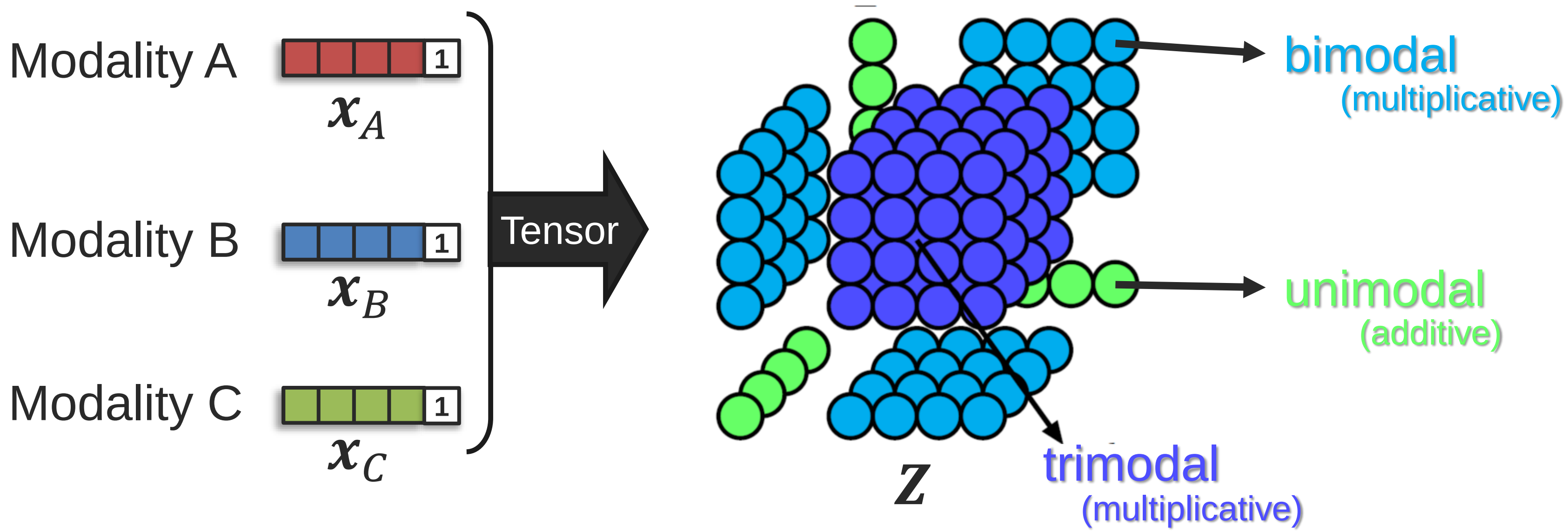
Jayakumar et al., Multiplicative Interactions and Where to Find Them, ICLR 2020

Tensor Fusion



Tensor Fusion (bimodal):

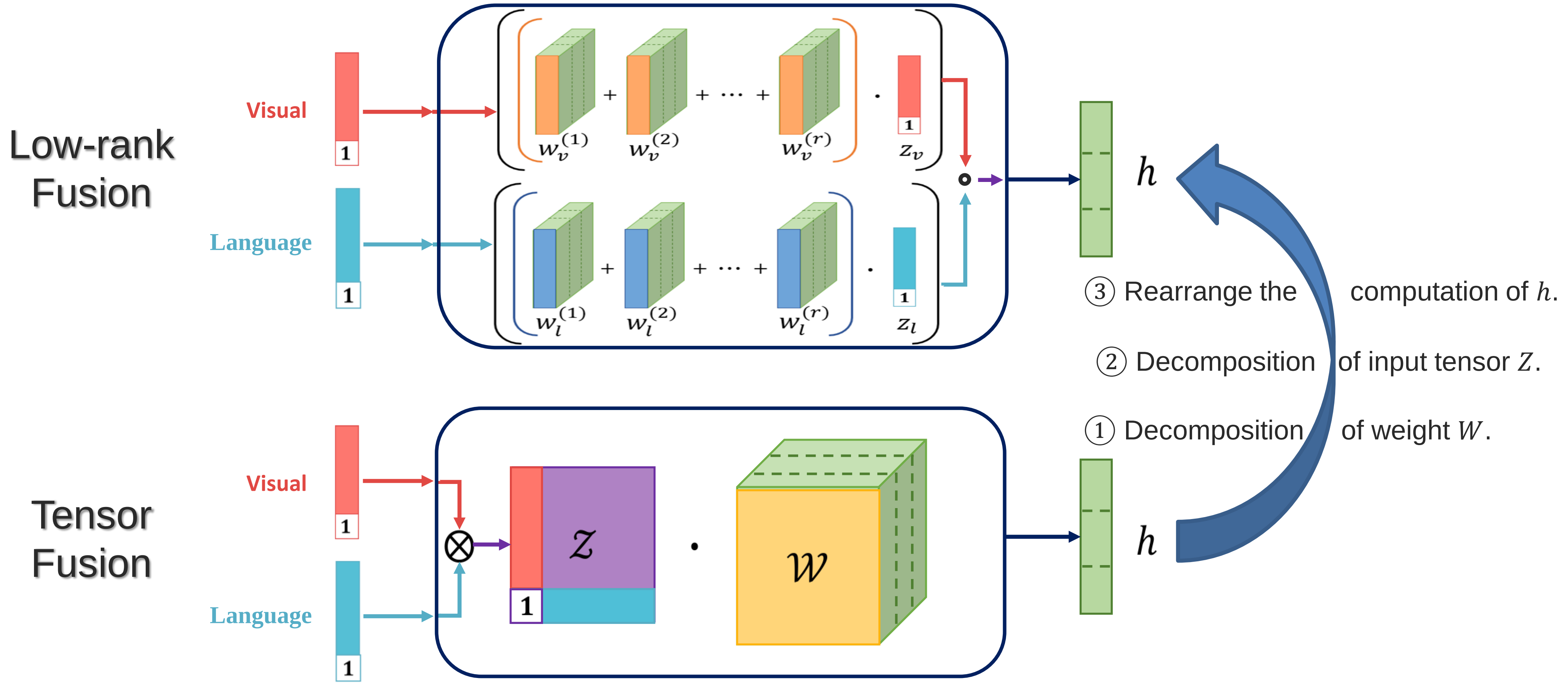
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

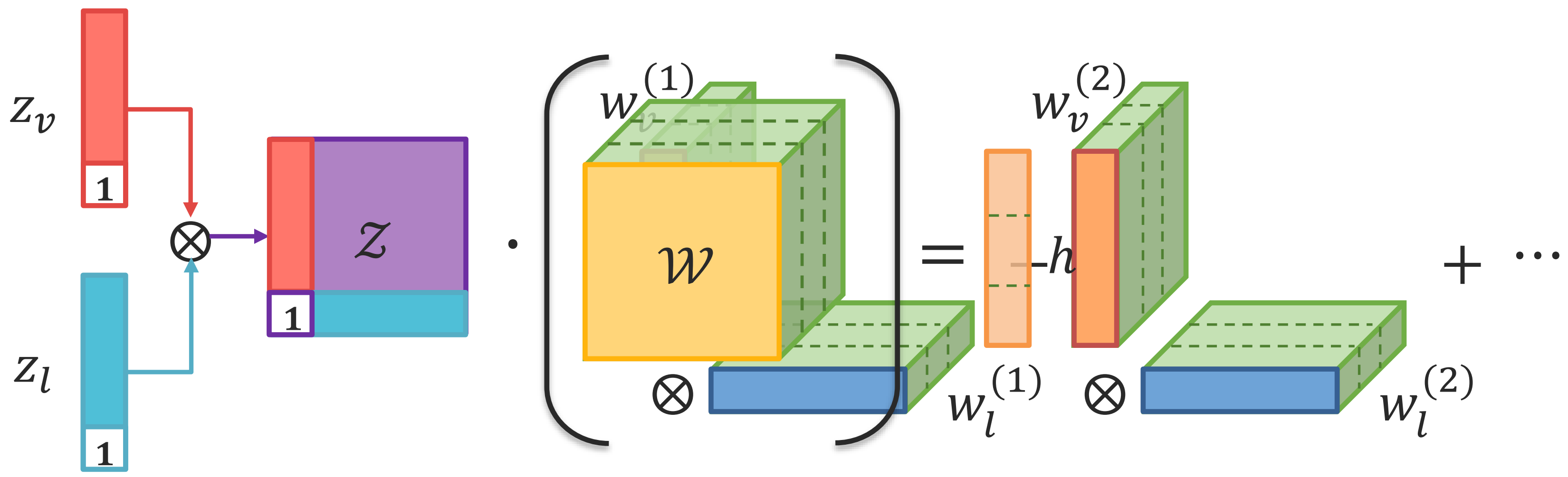
Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis, EMNLP 2017

Low-rank Fusion



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

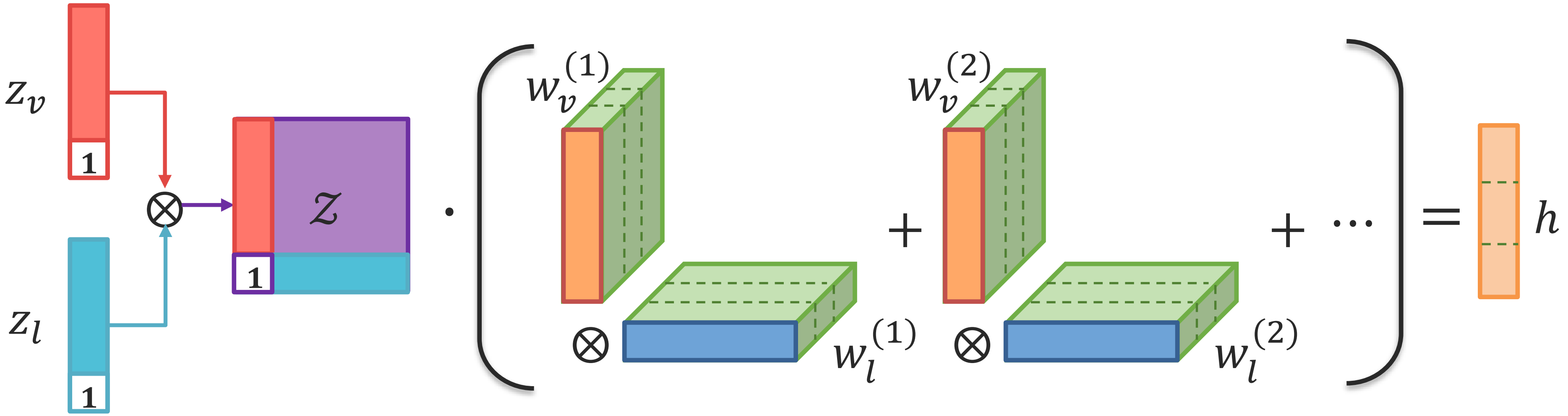
Low-rank Fusion



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

33

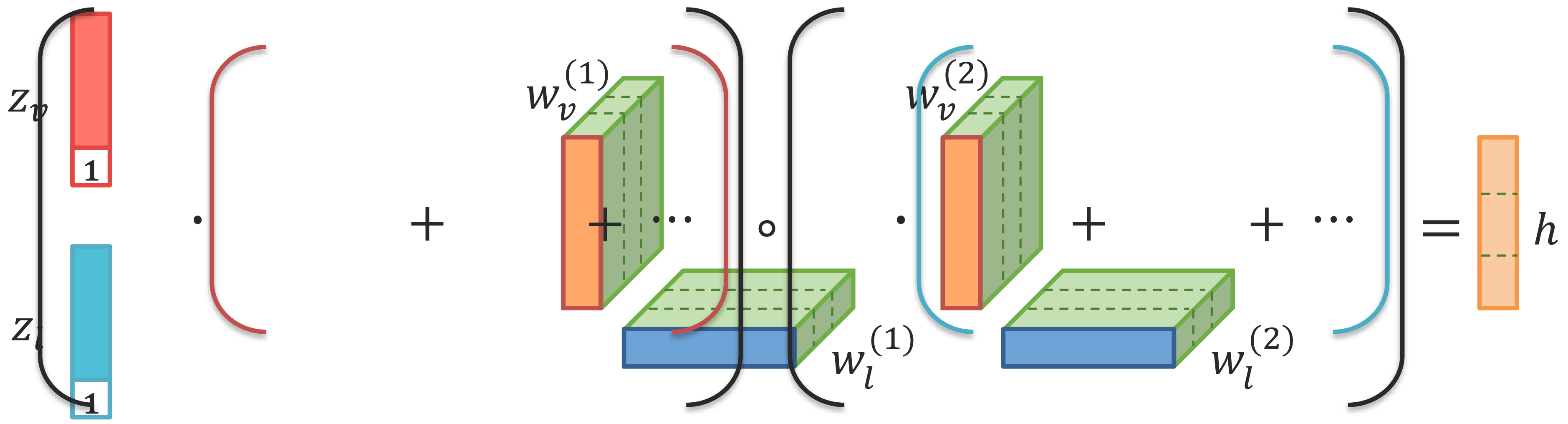
Low-rank Fusion



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

34

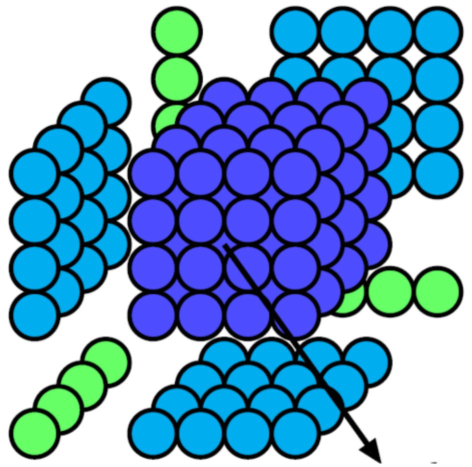
Low-rank Fusion



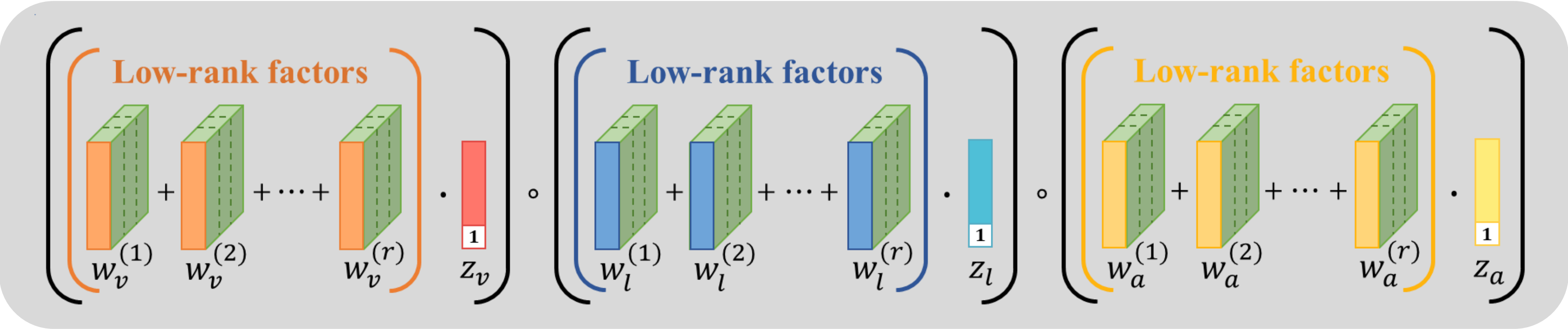
Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

Low-rank Fusion with Trimodal Input

Tensor Fusion



Low-rank Fusion :



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

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?

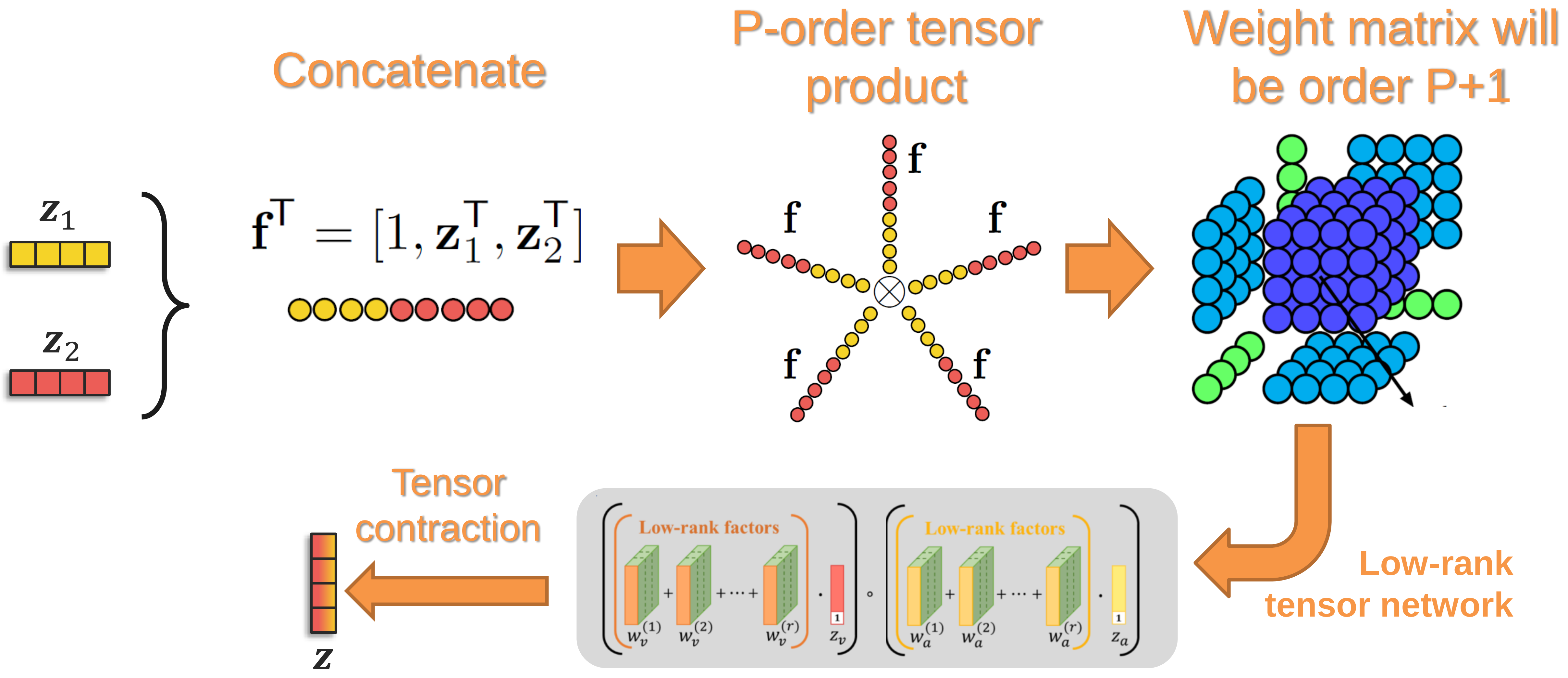
For example:

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

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

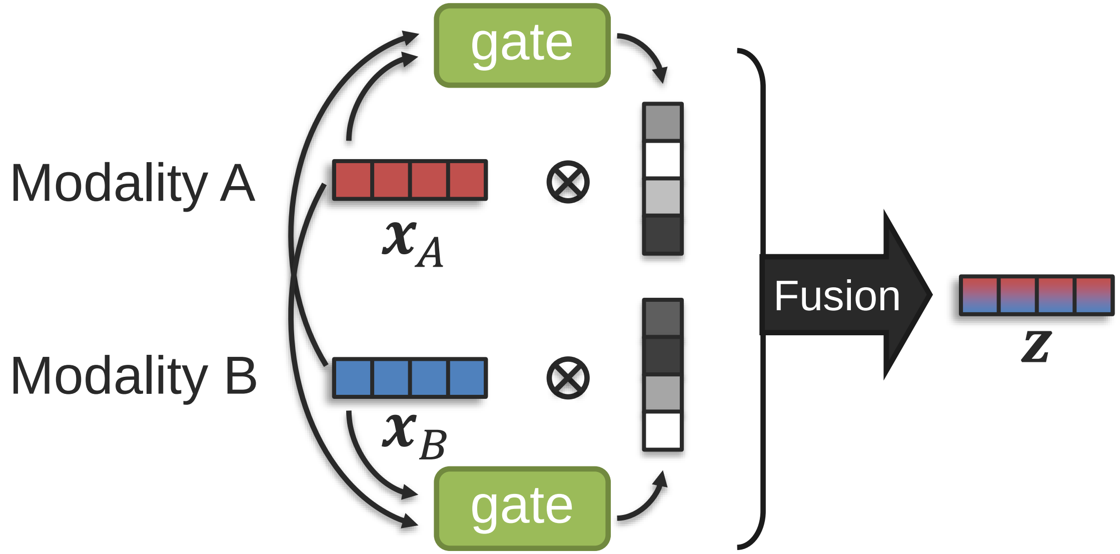
$$+w_{10} (x_B^3 \times x_C^3)$$

High-Order Polynomial Fusion



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

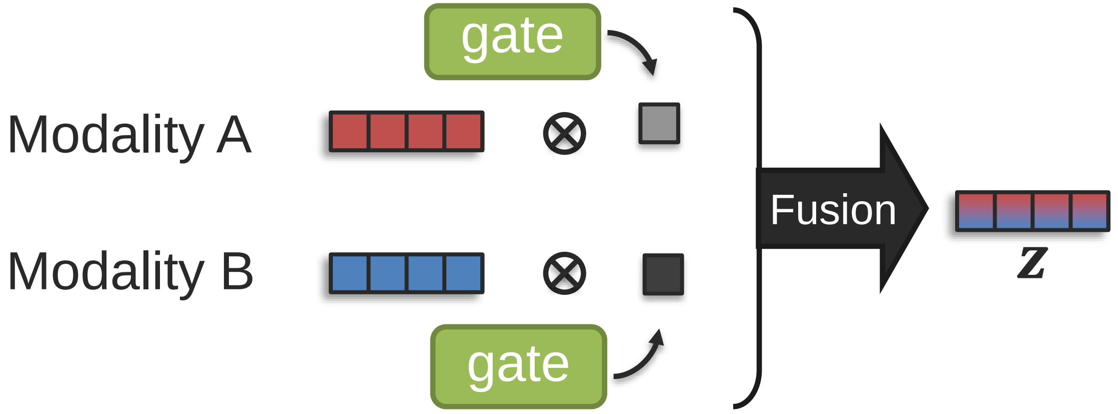
Gated Fusion



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



➡ Gating output can be one weight for the whole modality

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

Gating Module (aka, attention module)

Input → gate →

What should it be?

Target modality

Other modality

All modality

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

…or with a more positive view:

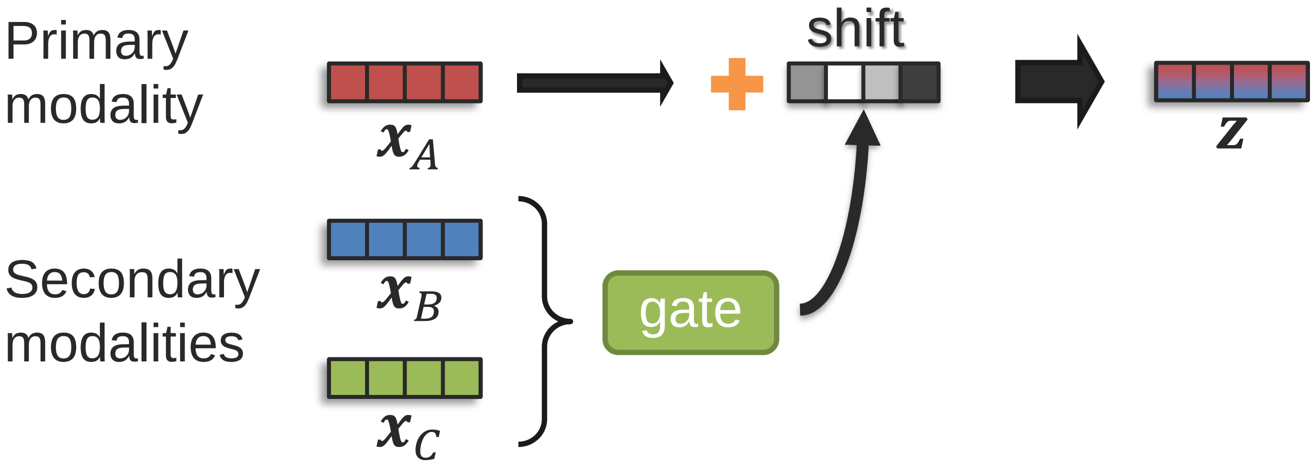
"Neural network designed to select preferable signal to move forward" (attention)

Soft attention gate → Easier to compute derivative (gradient)

Hard attention gate → Derivative is harder (e.g., use reinforcement learning)

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

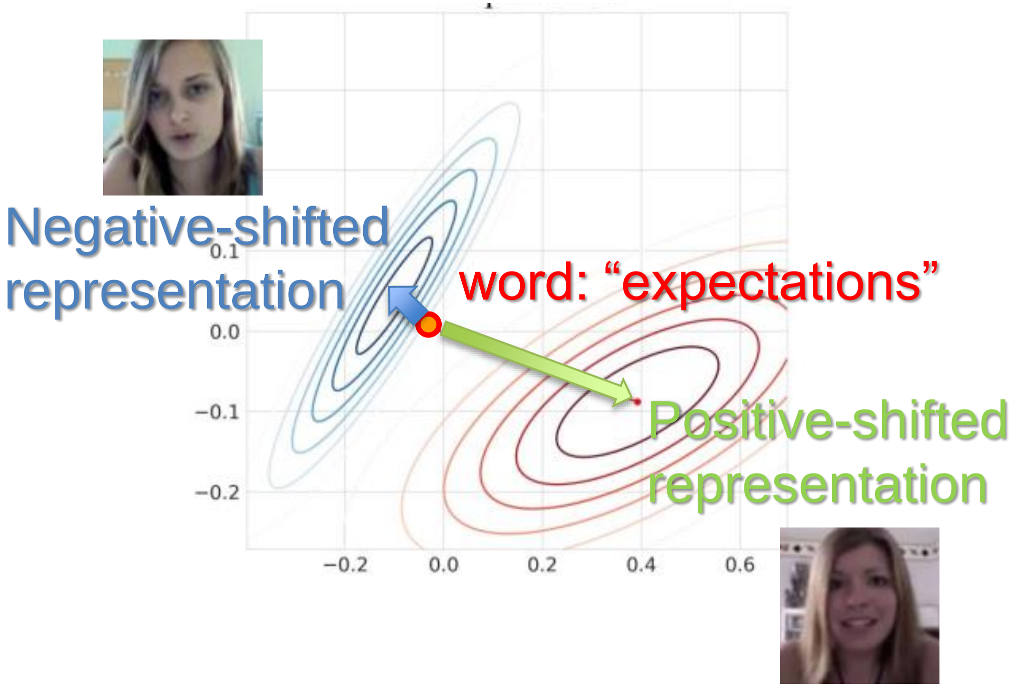
Modality-Shifting Fusion



Example with language modality:

Primary modality: language

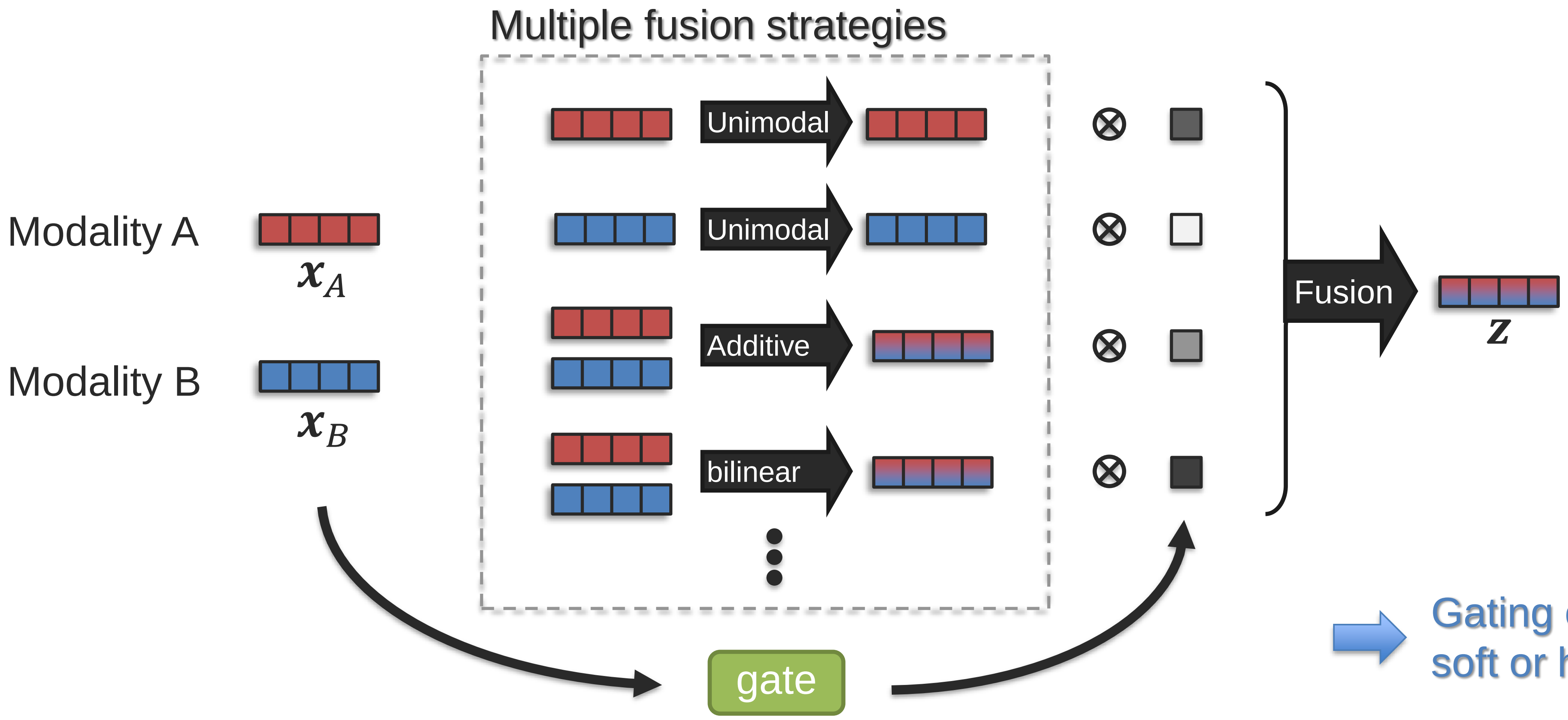
Secondary modalities: acoustic and visual



Wang et al., Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors, AAAI 2019

Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers, ACL 2020

Dynamic Fusion

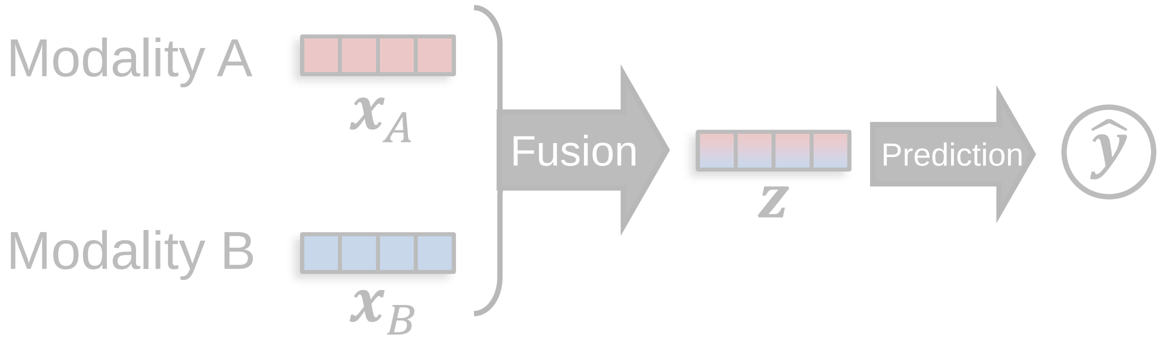


Gating can we with soft or hard attention

Zadeh et al., Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph, ACL 2018

Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records, AAAI 2021

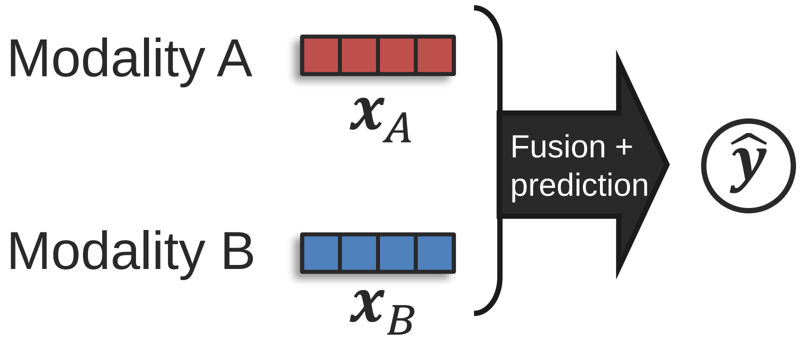
Nonlinear Fusion



Nonlinear fusion:

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

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

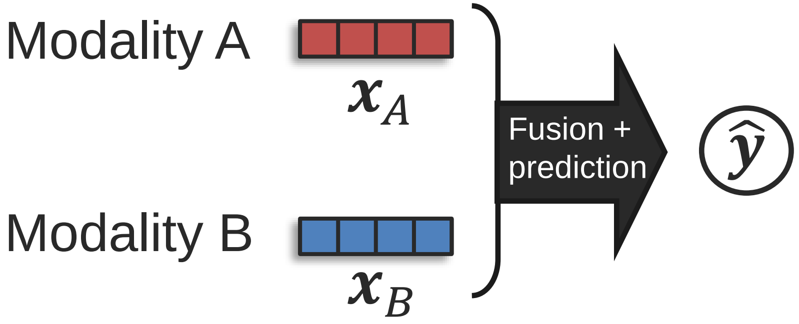


This could be seen as *early fusion*:

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

… but will our neural network learn the nonlinear interactions?

Measuring Non-Additive Interactions



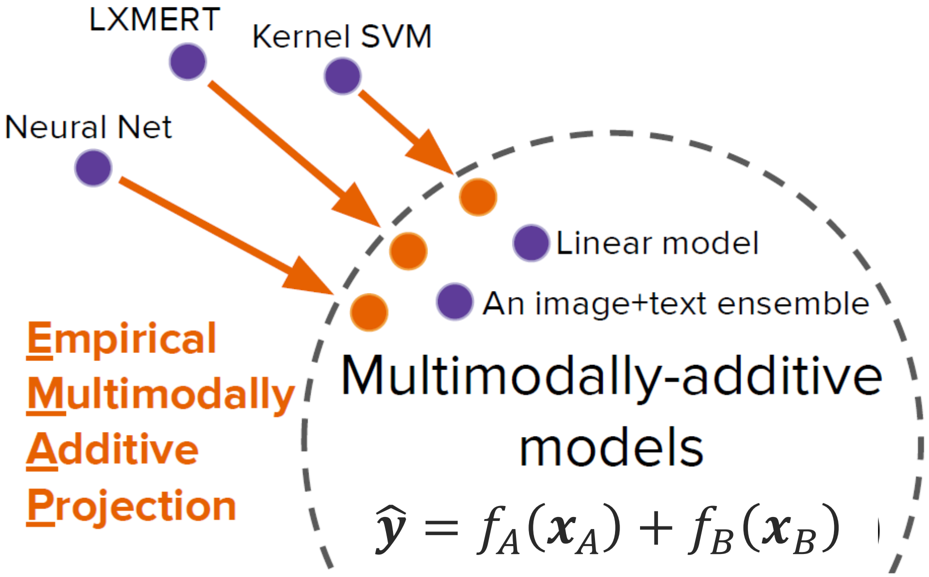
Nonlinear fusion:

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

Projection?

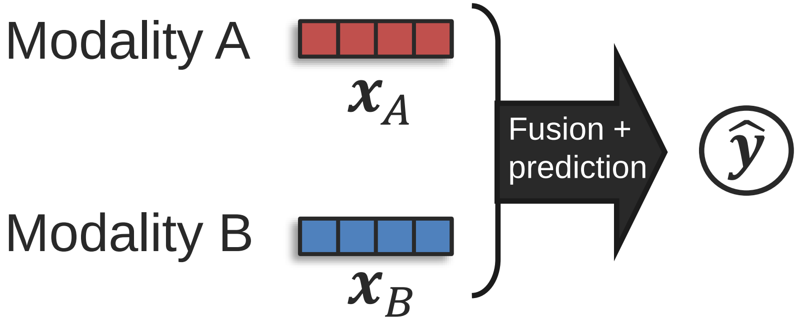
Additive fusion:

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



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

Measuring Non-Additive Interactions



Nonlinear fusion:

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

Projection?

Additive fusion:

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

Projection from nonlinear to additive (using EMAP):

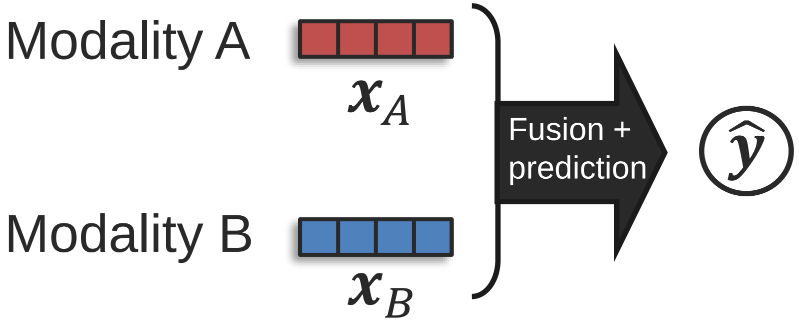
$$\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B) = \underbrace{\mathbb{E}_{\boldsymbol{x}_B}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_A(\boldsymbol{x}_A)} + \underbrace{\mathbb{E}_{\boldsymbol{x}_A}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_B(\boldsymbol{x}_B)} - \underbrace{\mathbb{E}_{\boldsymbol{x}_A,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

The expectations $\mathbb{E}$ can be approximated with summation over training data:

$$\hat{f}_A(\boldsymbol{x}_A) = \frac{1}{N}\sum_{j=1}^{N} f(\boldsymbol{x}_{A,j}, \boldsymbol{x}_{B,j})$$

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

Measuring Non-Additive Interactions



Nonlinear fusion:

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

EMAP projection

Additive fusion:

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

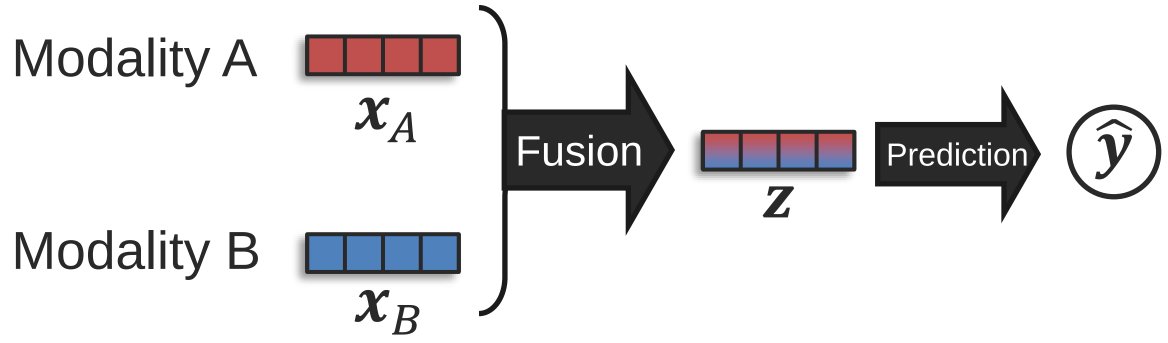
		I-INT	I-SEM	I-CTX	T-VIS	R-POP	T-ST1	T-ST2
Nonlinear	Neural Network	90.4	69.2	78.5	51.1	63.5	71.1	79.9
Polynomial	Polykernel SVM	**91.3**	**74.4**	**81.5**	50.8	–	72.1	**80.9**
Nonlinear	FT LXMERT	83.0	68.5	76.3	**53.0**	63.0	66.4	78.6
Nonlinear	↳ + Linear Logits	89.9	73.0	80.7	**53.4**	**64.1**	**75.5**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

Always a good baseline!

Differences are small!!!

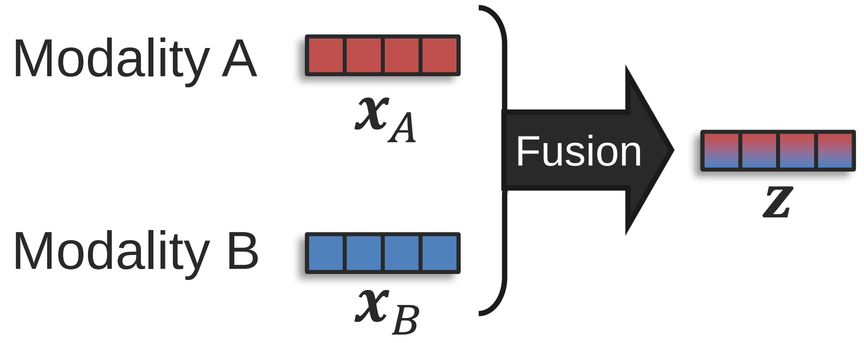
Hessel and Lee, Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!, EMNLP 2020 ⟶ introduced the EMAP method

Learning Fusion Representations



How to learn fusion models?

Learning Fusion Representations

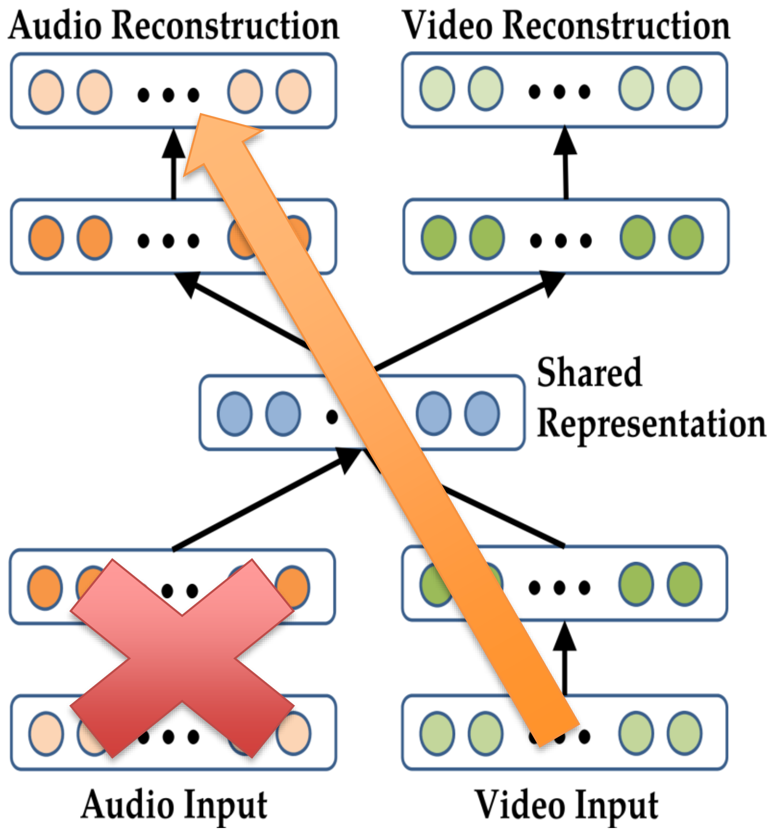


How to learn fusion models?

What will be the loss function?

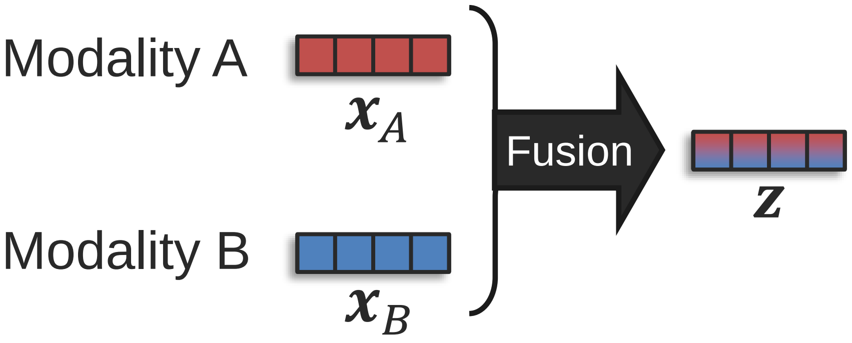
Can it hallucinate the other modality?

Multimodal Autoencoder

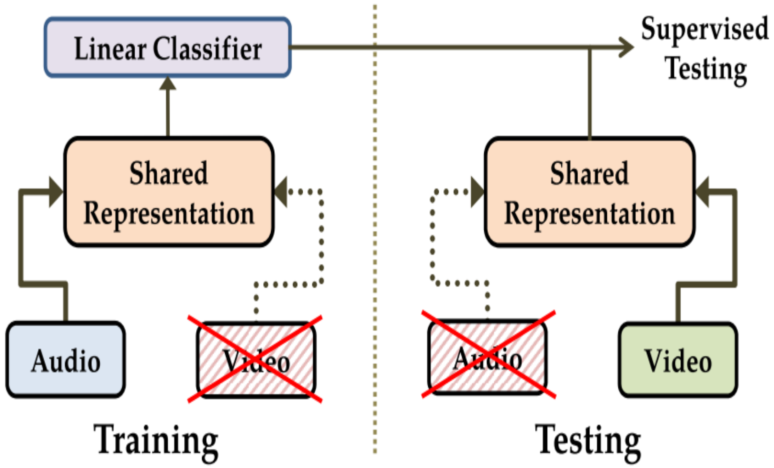


Ngiam et al, Multimodal Deep Learning, 2011

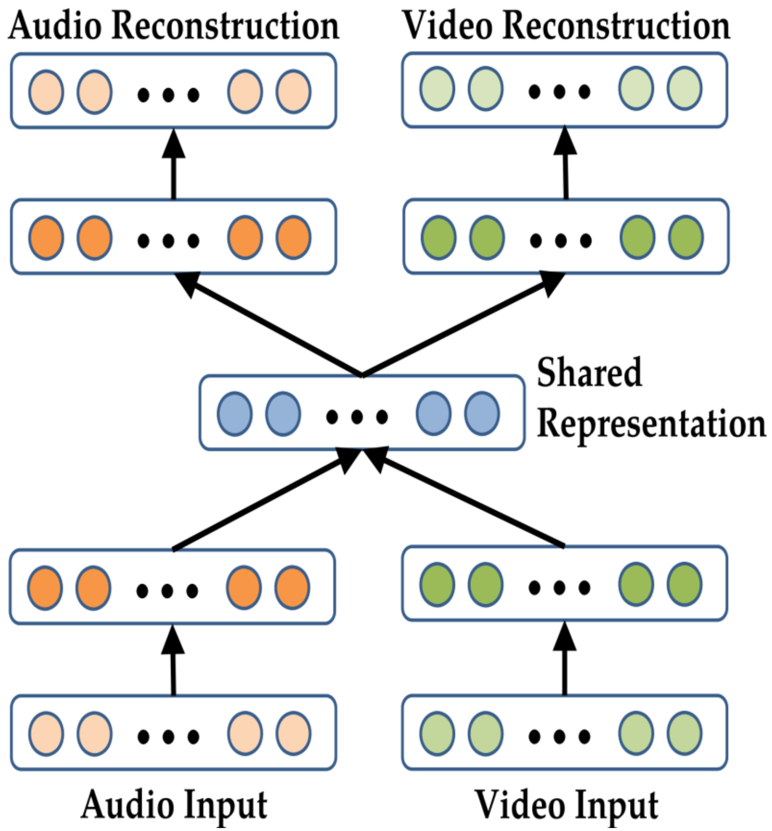
Learning Fusion Representations



Interesting experiment: "Hearing to see"

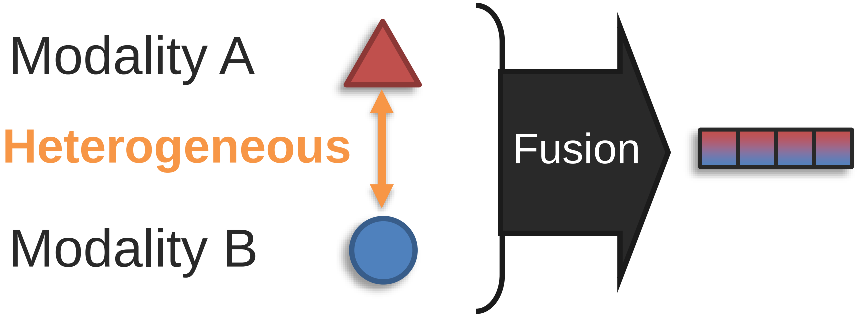


Multimodal Autoencoder



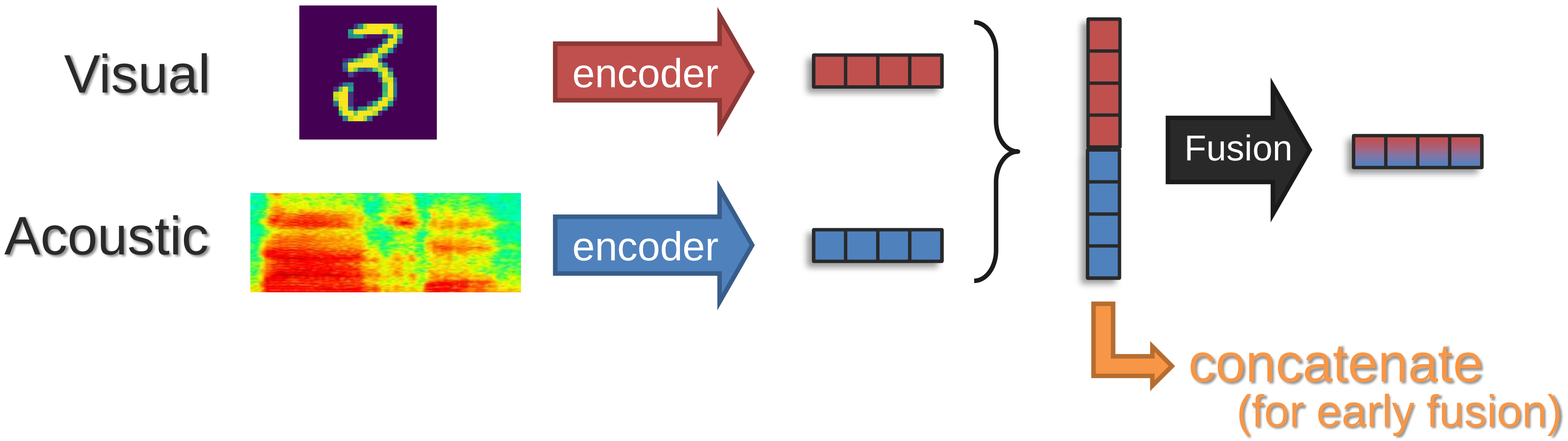
Ngiam et al, Multimodal Deep Learning, 2011

Fusion with Raw Modalities

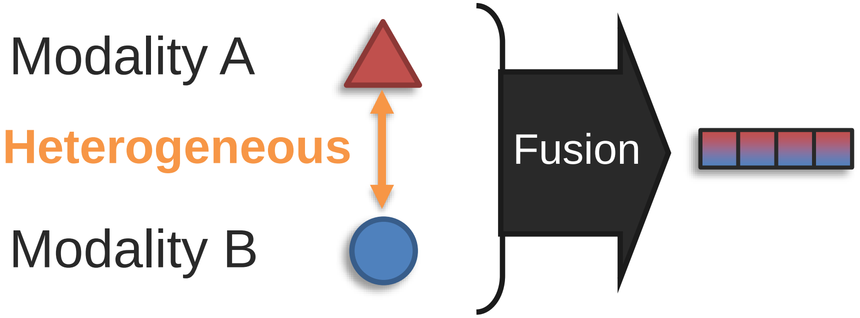


Open Challenge!

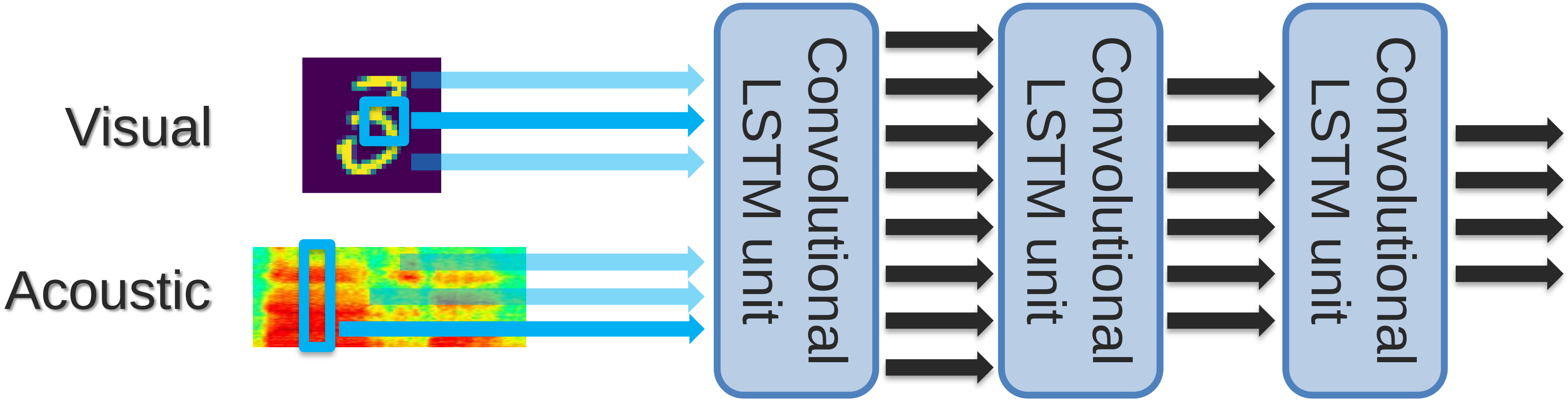
Example: From Early Fusion…



Fusion with Raw Modalities

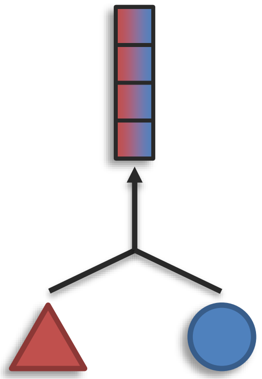


Example: From Early Fusion… to Very Early Fusion (inspired by human brain)



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

Sub-Challenge 1a: Representation Fusion



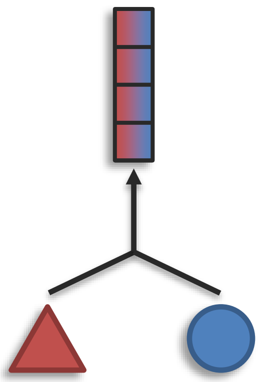
Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Challenge 1: Representation

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

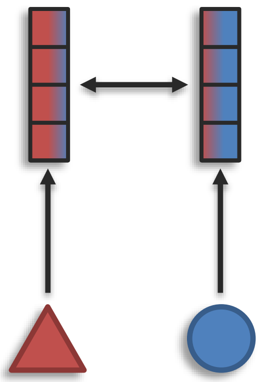
Sub-challenges:

Fusion



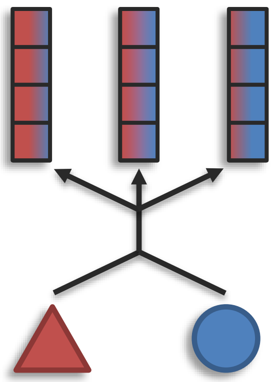
modalities > # representations

Coordination



modalities = # representations

Fission



modalities < # representations

Next week