Multimodal AI

Lecture 4.1 – Multimodal Fusion

Paul Liang
Assistant Professor
MIT Media Lab & MIT EECS

https://pliang279.github.io
ppliang@mit.edu
@pliang279

mit media lab

MIT EECS

Assignments for This Coming Week

Project proposal due today. Meet with me at 4-5pm if you want feedback.

I want to meet every group at least once regarding their project ideas either today or this Thursday.

Compute credits: 40 x $50 Kimi credits, 40 x $40 other credits.

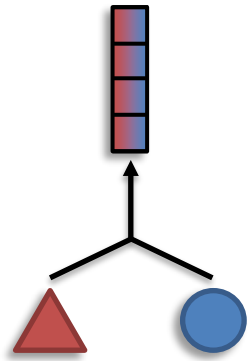
HW2 released last week, due next Wednesday 3/4.

Today's lecture

1. Basics of multimodal interactions and fusion
2. Early, intermediate, late fusion
3. Multiplicative and dynamic fusion
4. Quantifying fusion

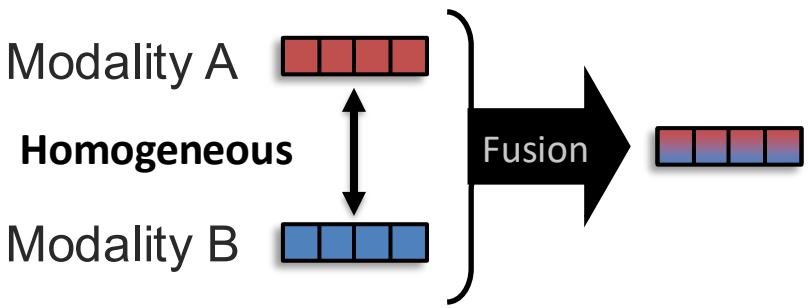
[credit: some slides in this lecture were co-developed with Louis-Philippe Morency for CMU course 11-777]

Multimodal Fusion

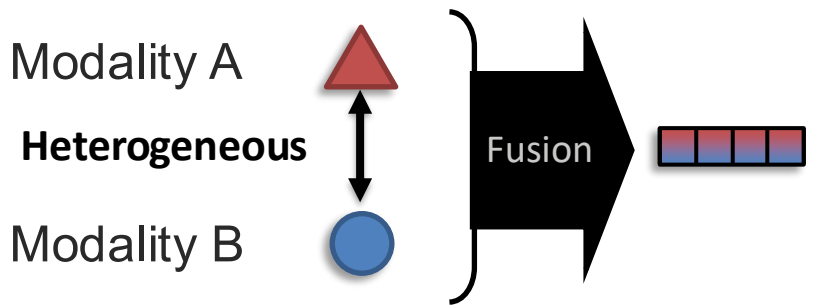


Definition: Learn a joint representation that models the **multimodal interactions** between individual elements of different modalities.

Fusion with abstract modalities:



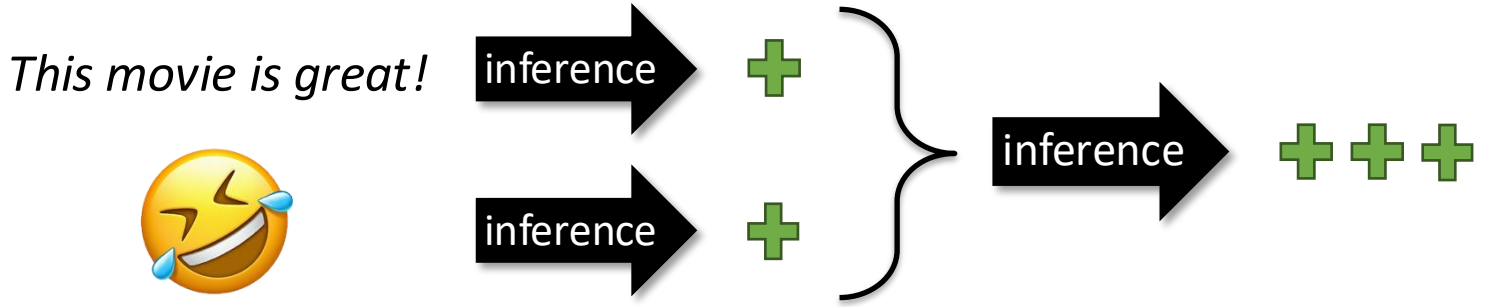
Fusion with raw modalities:

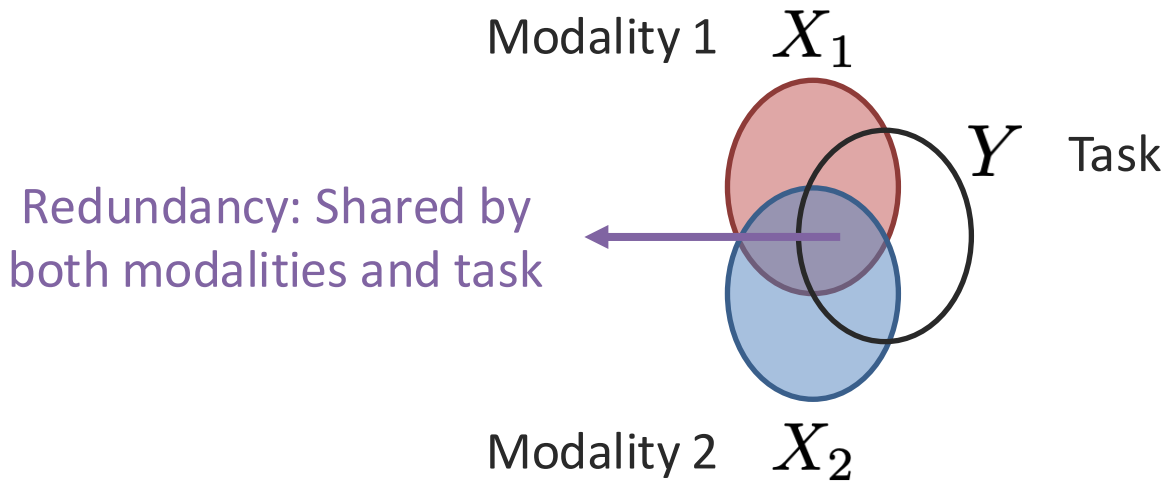


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Multimodal Interactions

Interactions: How modalities *combine* to provide information for a task.

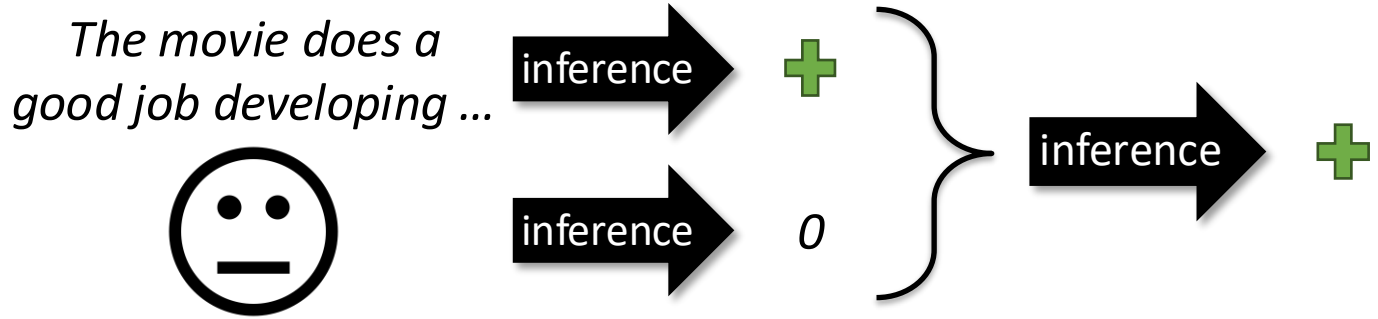


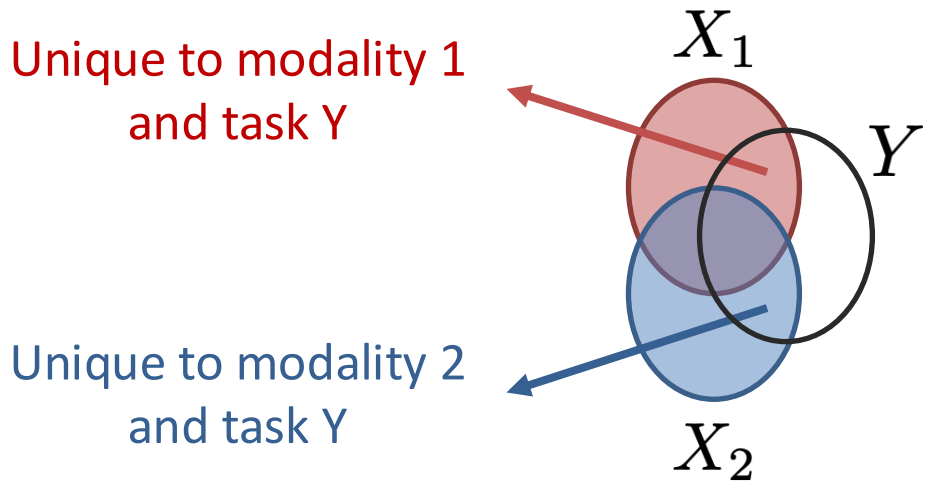


multisensory intelligence

Multimodal Interactions

Interactions: How modalities *combine* to provide information for a task.

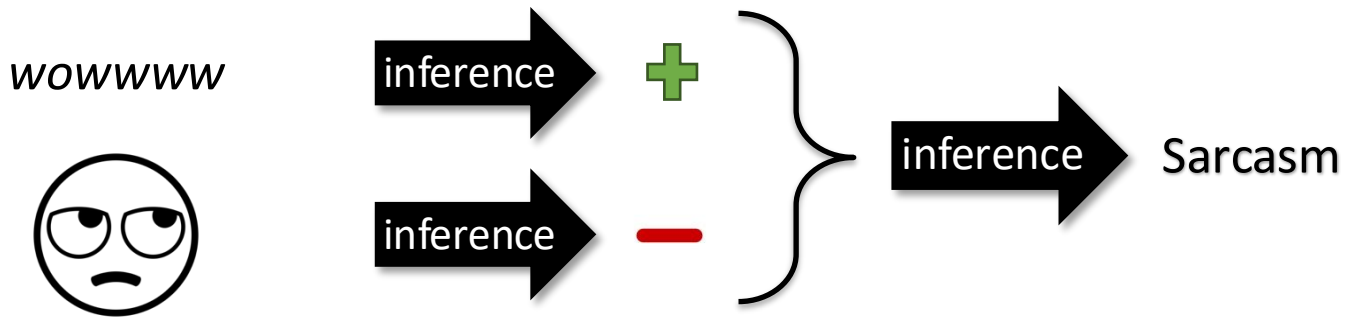


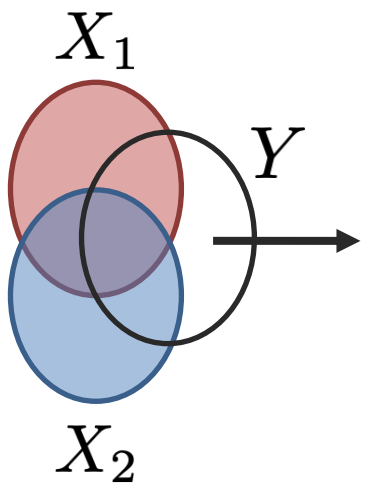


multisensory intelligence

Multimodal Interactions

Interactions: How modalities *combine* to provide information for a task.

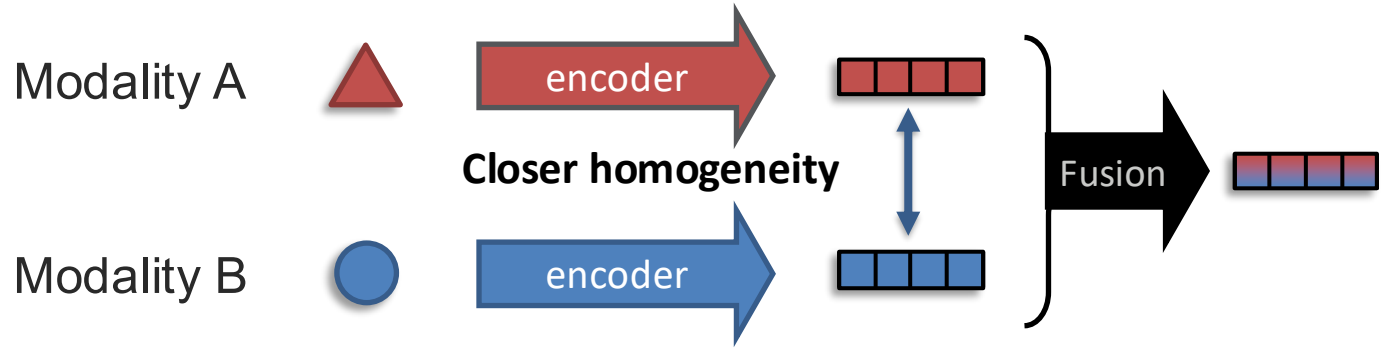




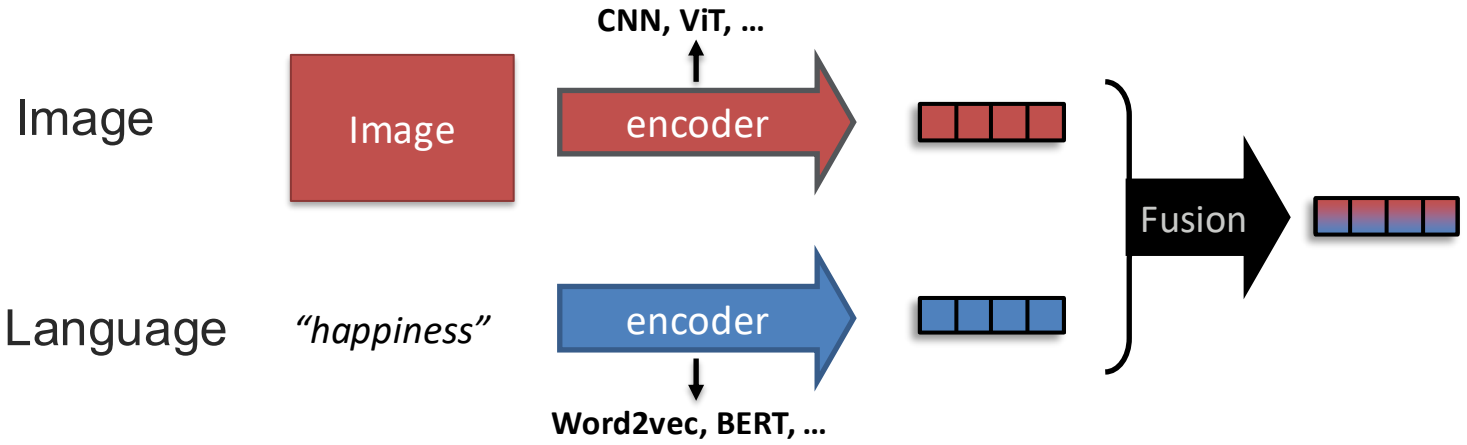
Synergy: Emerging information from multimodal interaction

multisensory intelligence

Fusion with Abstract Modalities



Example:



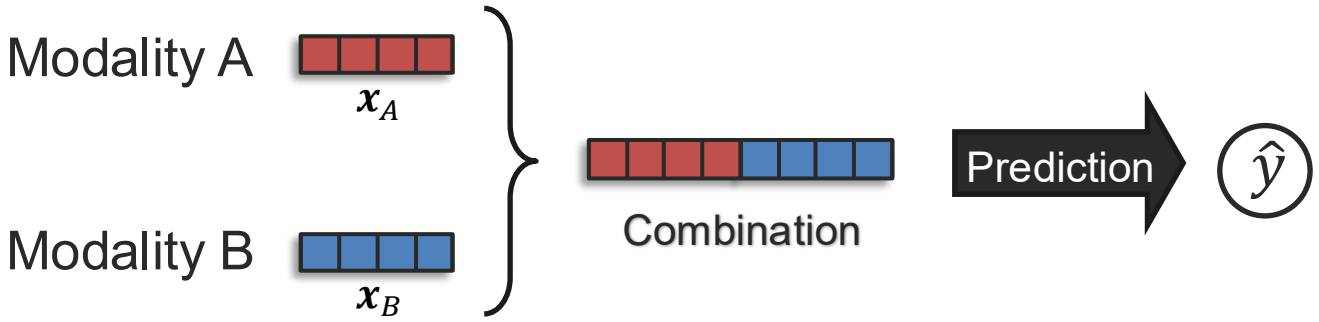
➡ Unimodal encoders can be jointly learned with fusion network, or pre-trained

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

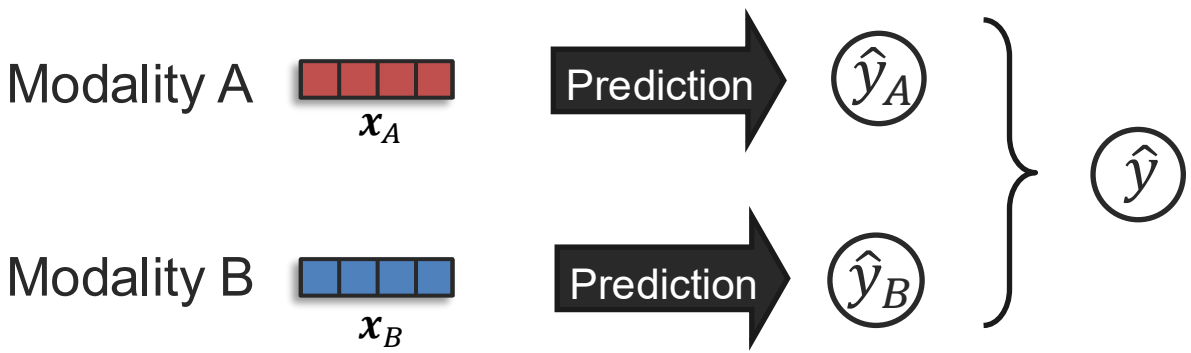
multisensory intelligence

Early and Late Fusion

Early fusion:



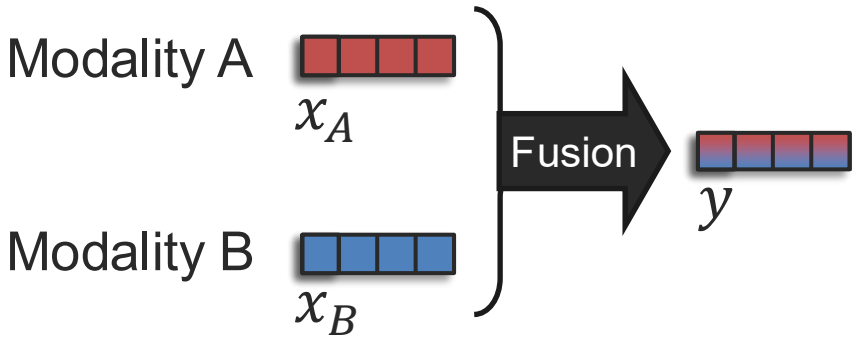
Late fusion:



[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Basic Concepts for Fusion



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

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

Linear regression:

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

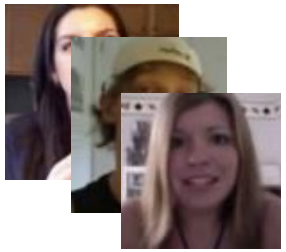
- w_0: intercept (bias term)
- $w_1 x_A + w_2 x_B$: Additive terms
- $w_3 (x_A \times x_b)$: Multiplicative term
- ϵ: error (residual term)

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

multisensory intelligence

Linear Fusion Case

300 book reviews



y: audience score

x_A: percentage of smiling

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

Linear regression:

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

- intercept (bias term)
- Additive terms
- Multiplicative term
- 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

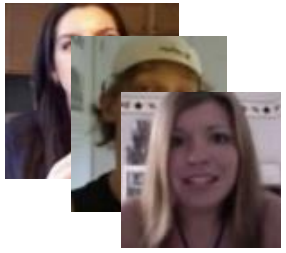
w_3: effect from x_A and x_B interaction only

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

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Linear Fusion Case

300 book reviews



y: audience score

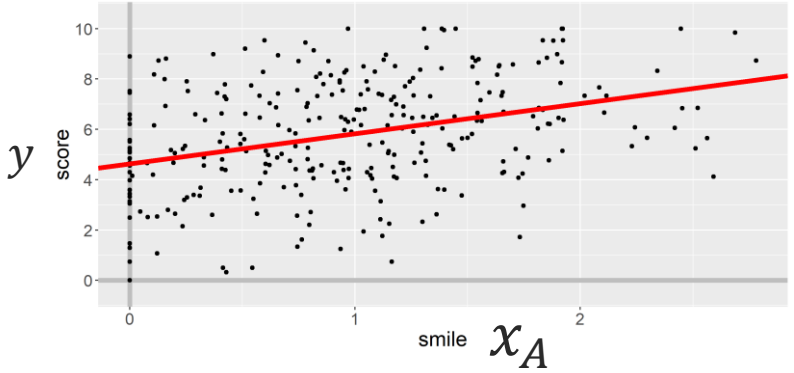
x_A: percentage of smiling

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

Linear regression:

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

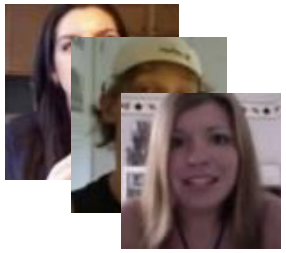
slope



	Estimate
w_0	4.63
w_1	1.20

Linear Fusion Case

300 book reviews



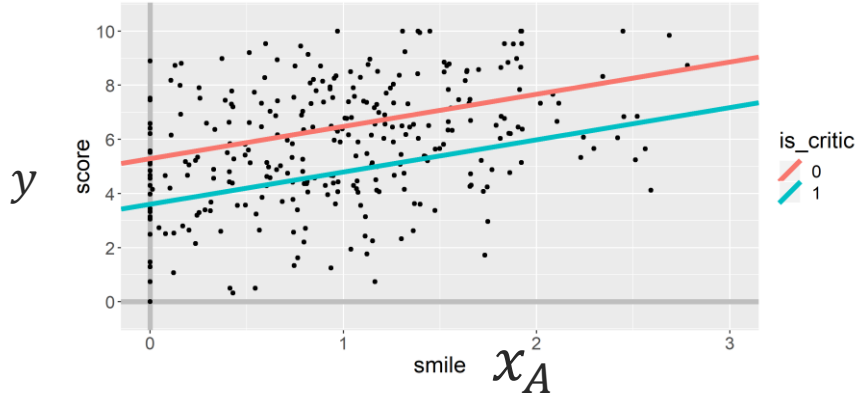
y: audience score

x_A: percentage of smiling

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

Linear regression:

$$y = w_0 + w_1 x_A + w_2 x_B + \epsilon$$



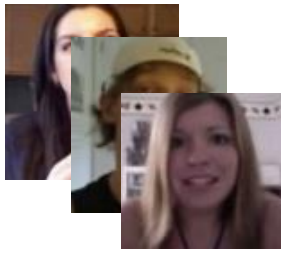
	Estimate
w_0	5.29
w_1	1.19
w_2	−1.69

Positive effect

Negative effect

Linear Fusion Case

300 book reviews



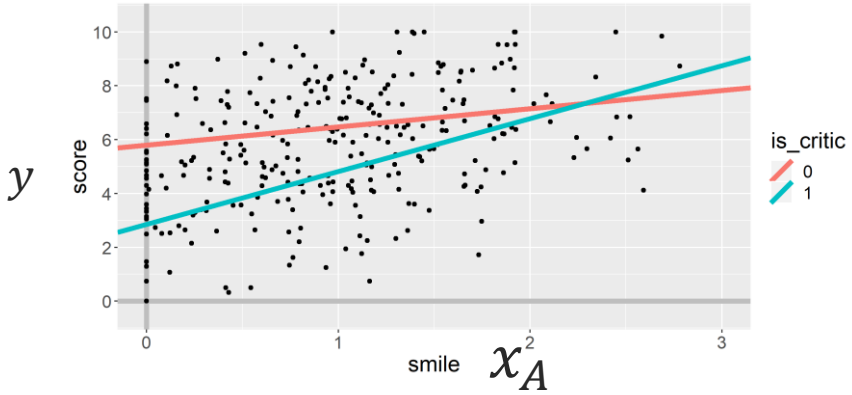
y: audience score

x_A: percentage of smiling

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

Linear regression:

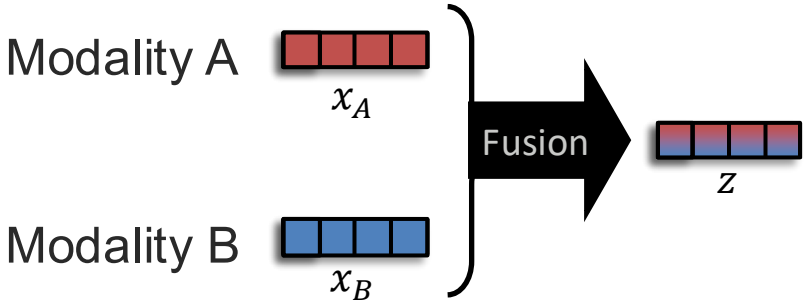
$$y = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_B) + \epsilon$$



	Estimate
w_0	5.79
w_1	0.68
w_2	−2.94
w_3	1.29

→ **Multiplicative interaction!**

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:

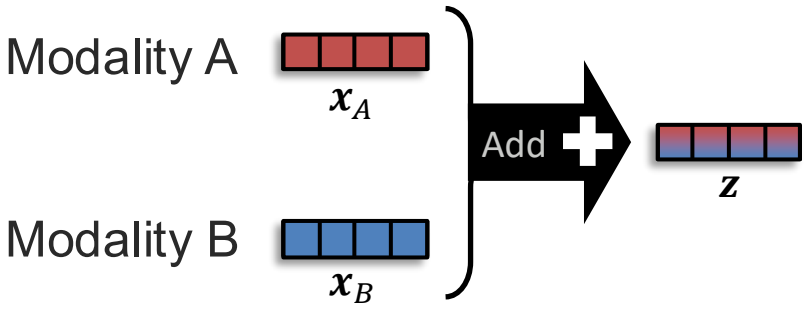
$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Additive Fusion

→ Back to multivariate case!

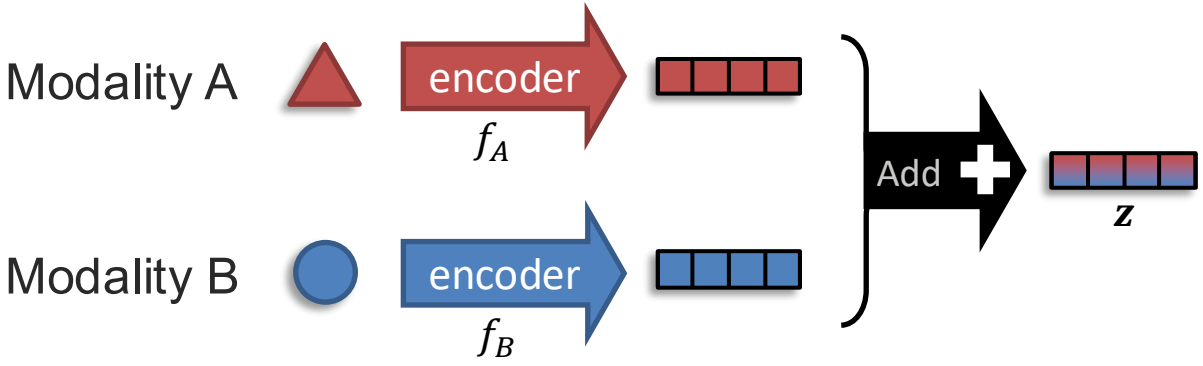
↳ (multi-dimensional features)



Additive fusion:

$$z = w_1 x_A + w_2 x_B$$

With unimodal encoders:



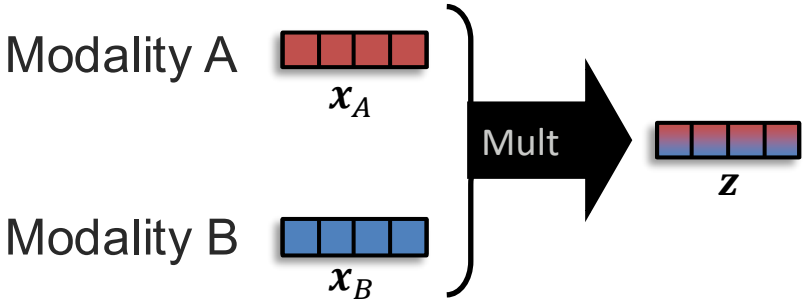
Additive fusion:

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

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

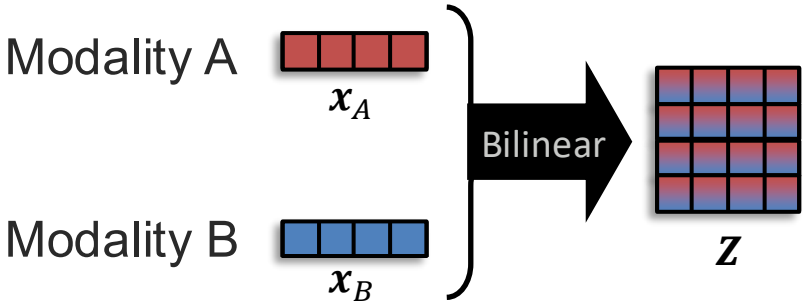
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Multiplicative Fusion



Multiplicative fusion:

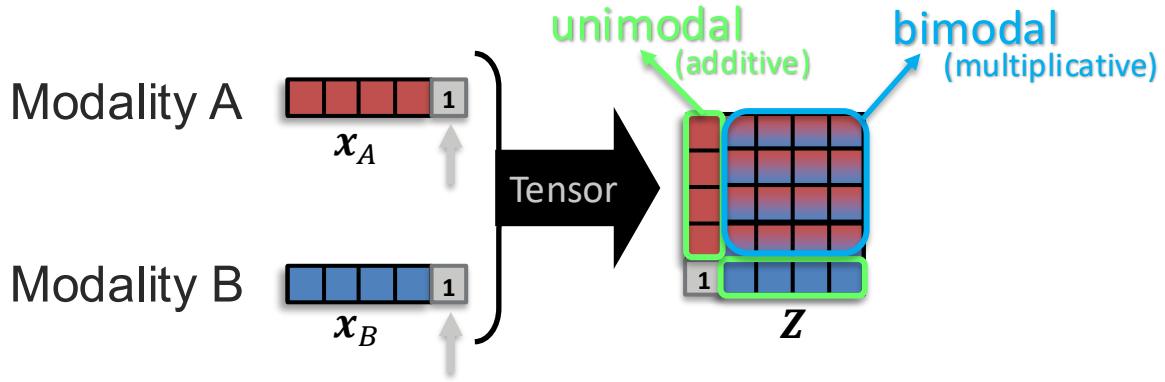
$$z = w(x_A \times x_B)$$



Bilinear Fusion:

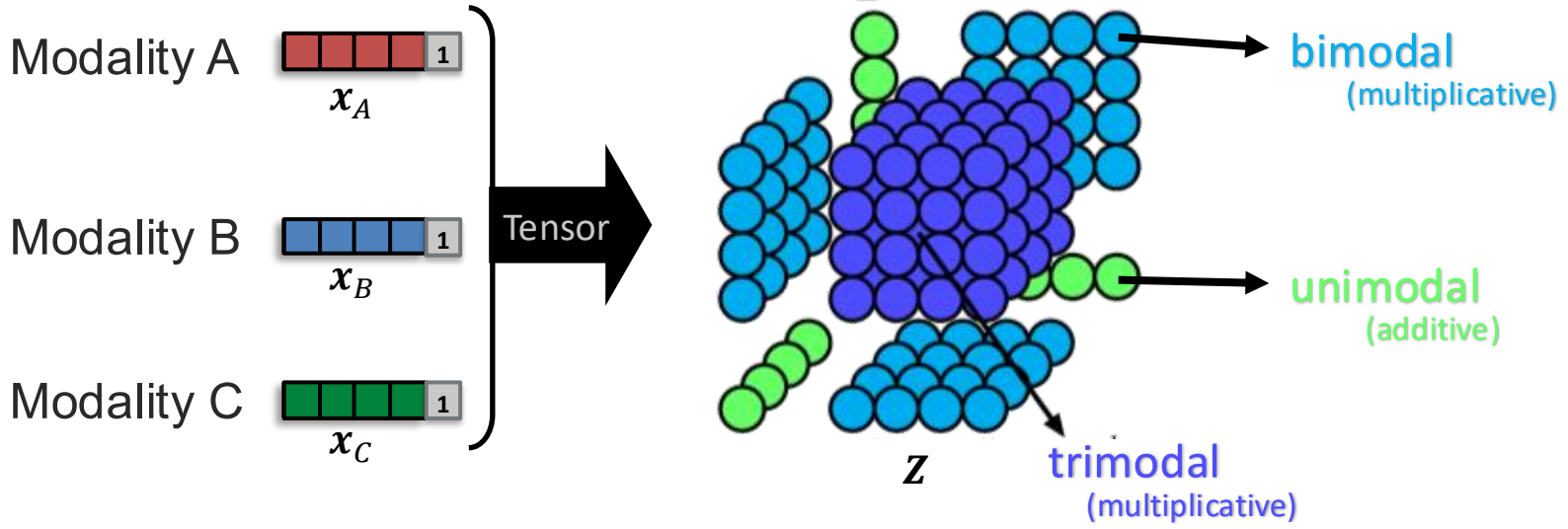
$$Z = w(x_A^T x_B)$$

Tensor Fusion



Tensor Fusion (bimodal):

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

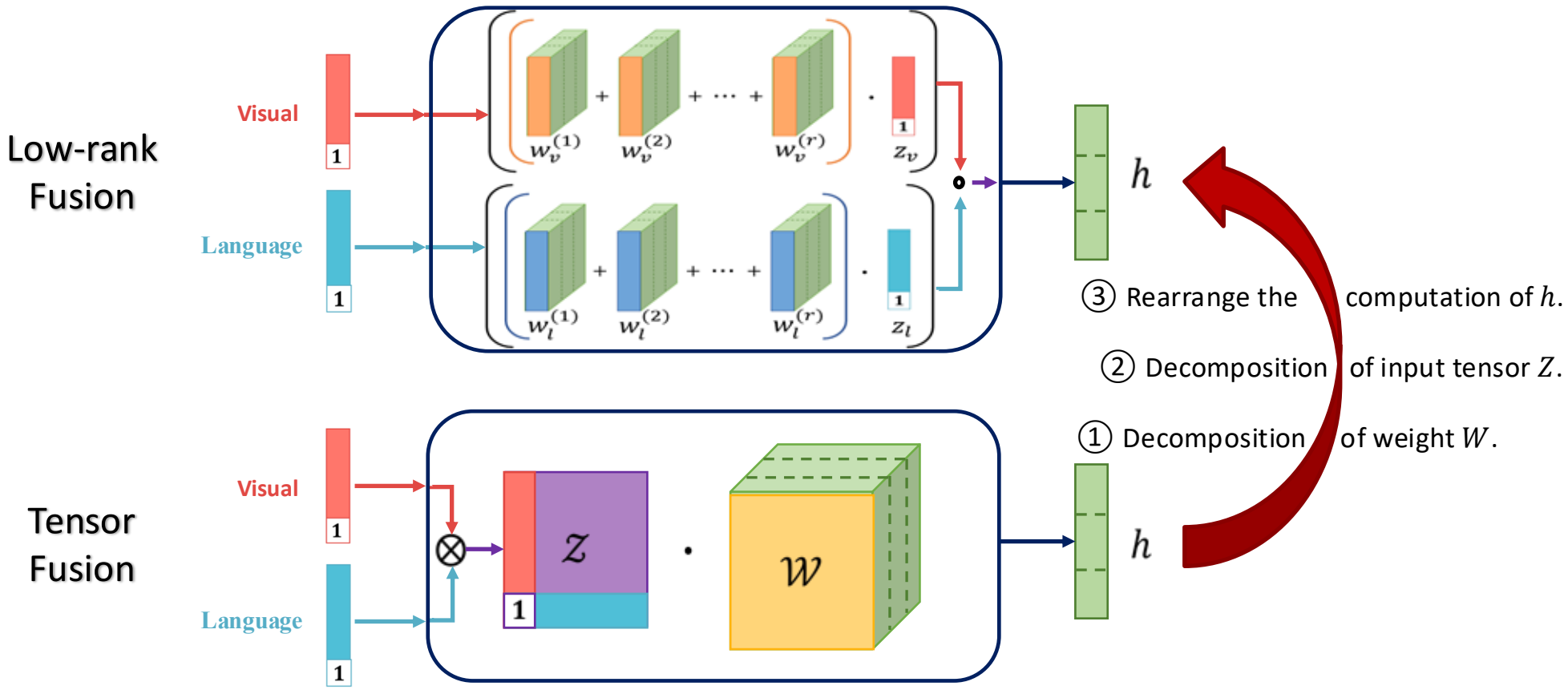


… but the weight matrix may end up quite large!

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

multisensory intelligence

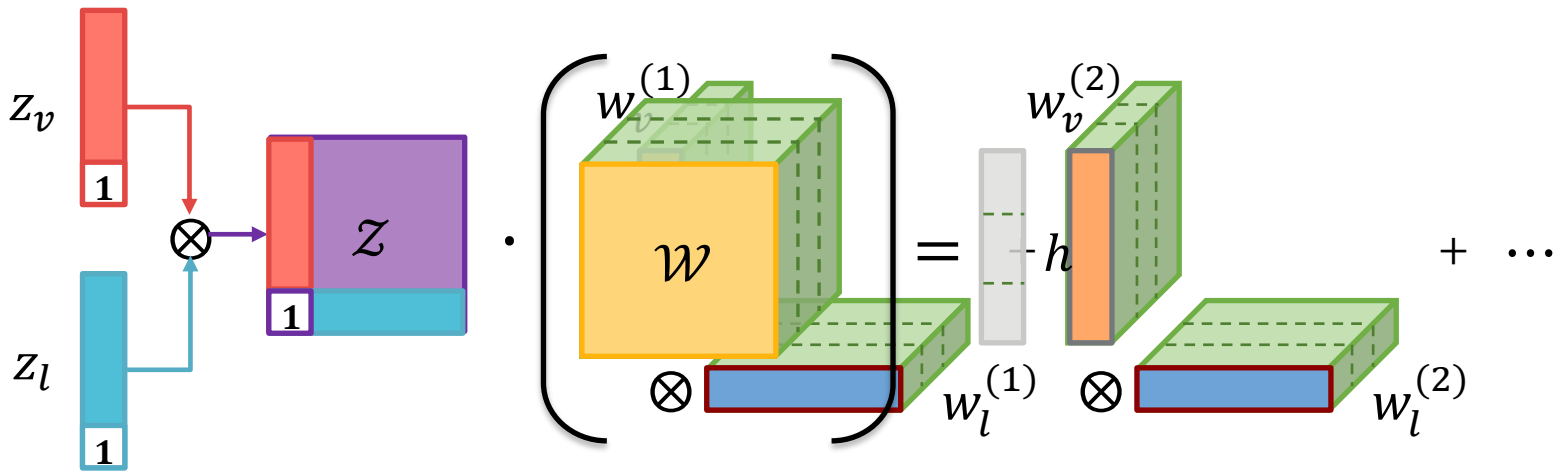
Low-rank Fusion



[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]
[Hu et al., LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022]

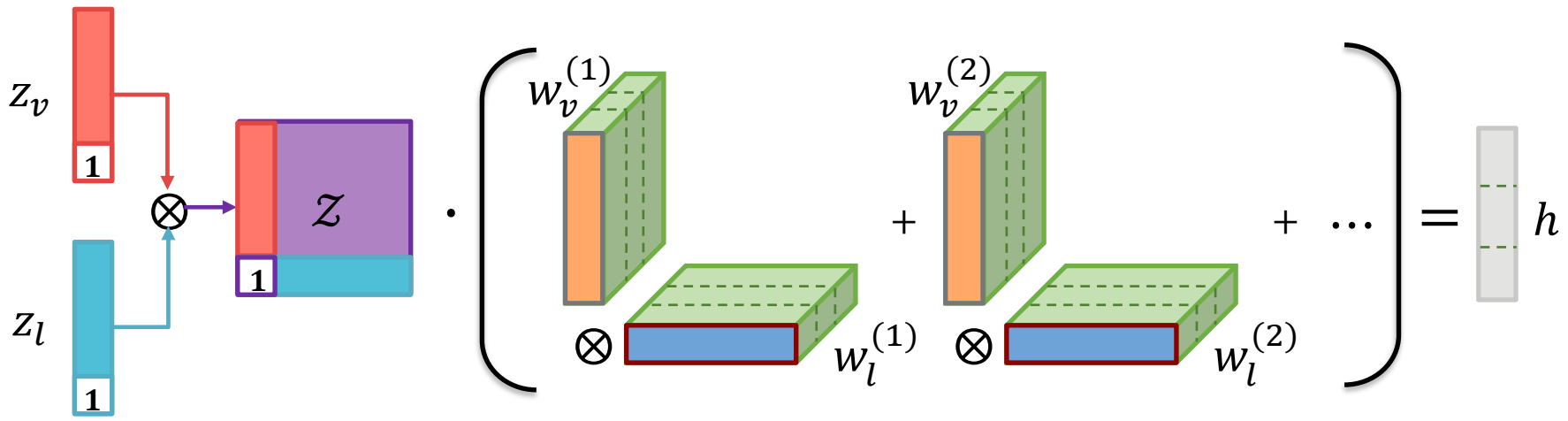
multisensory intelligence

Low-rank Fusion



[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]
[Hu et al., LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022]

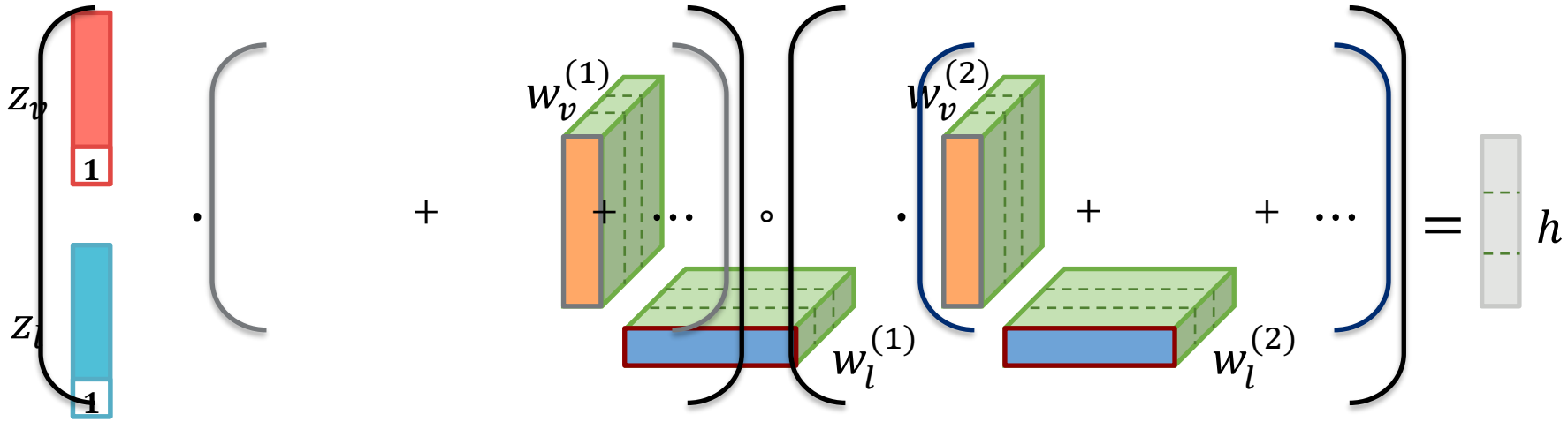
Low-rank Fusion



[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]
[Hu et al., LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022]

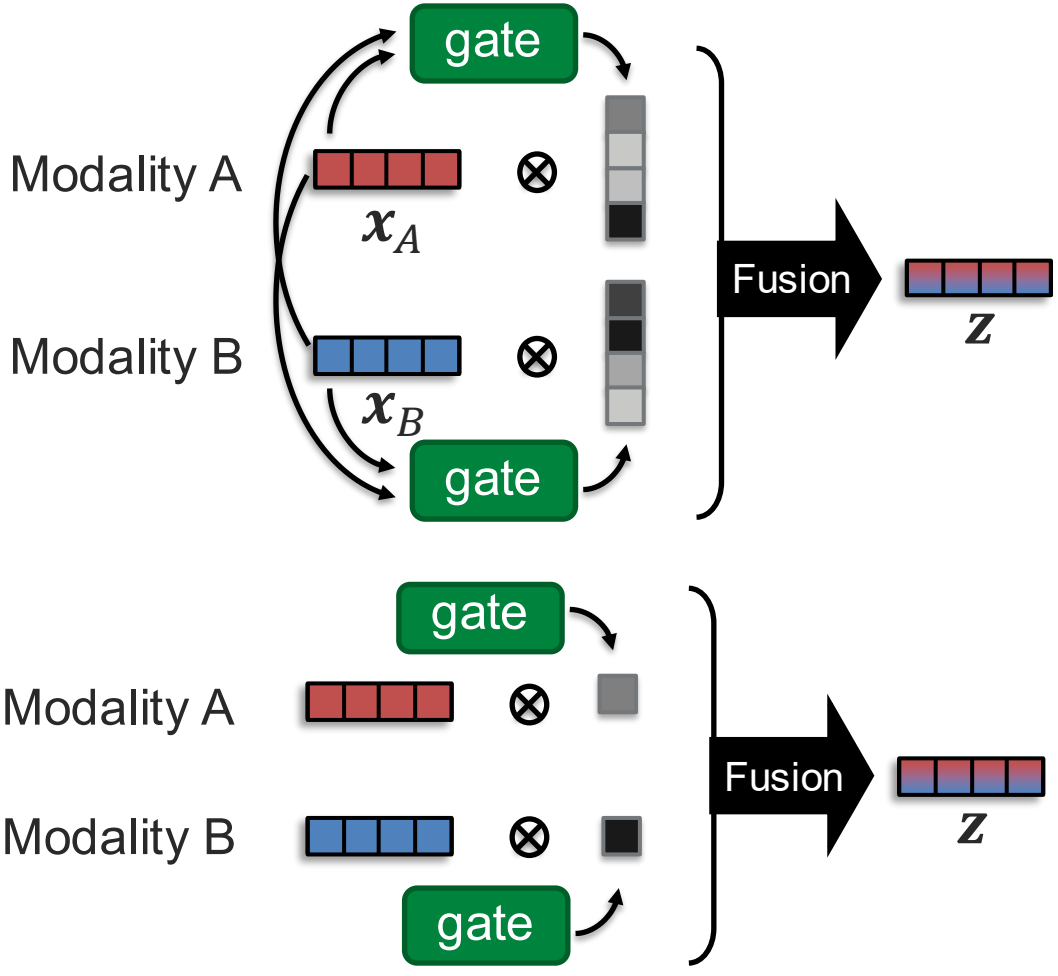
multisensory intelligence

Low-rank Fusion



[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]
[Hu et al., LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022]

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]

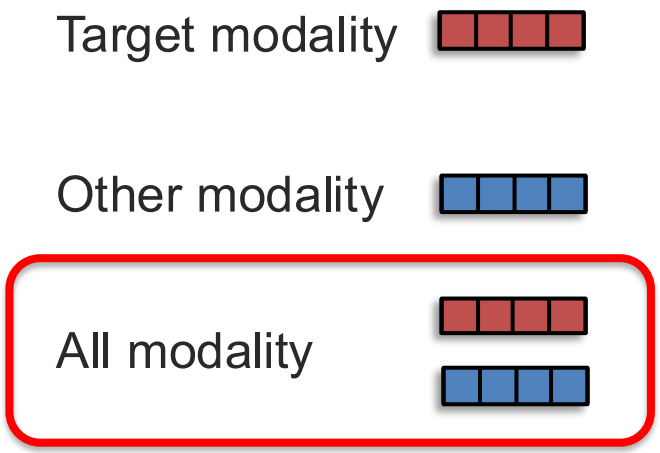
multisensory intelligence

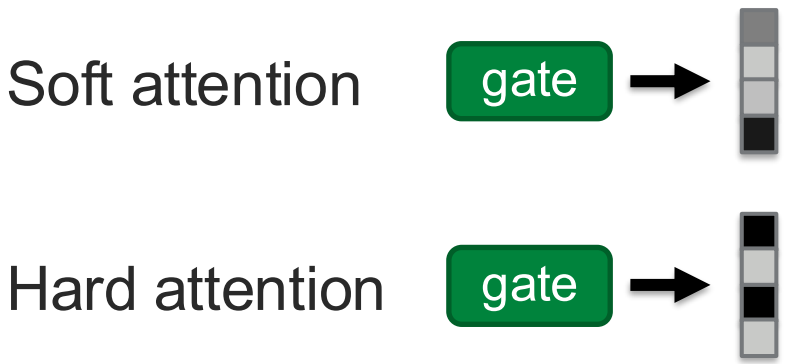
Gated Fusion

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

…or with a more positive view:

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

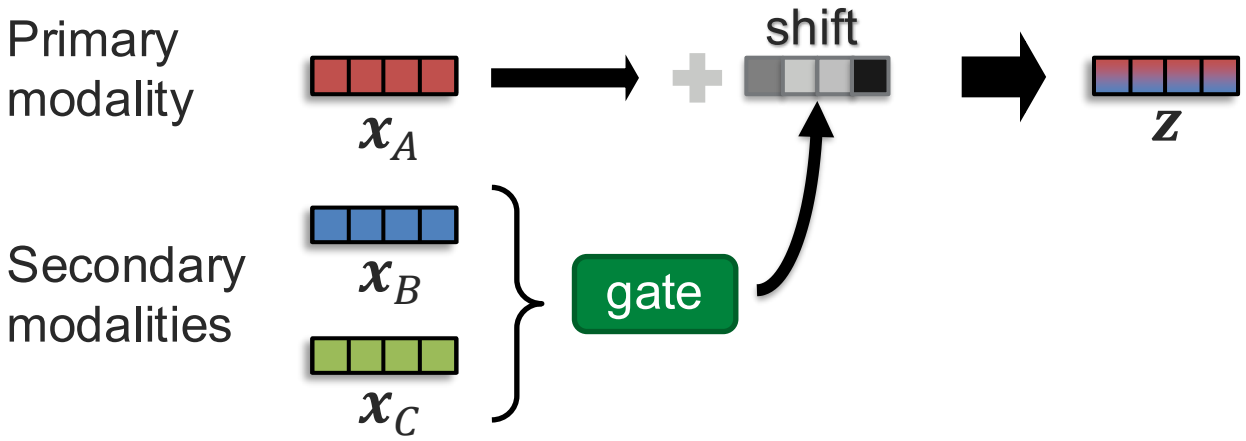




[Chen et al., Multimodal Sentiment Analysis with Word-level Fusion and Reinforcement Learning, ICMI 2017]

multisensory intelligence

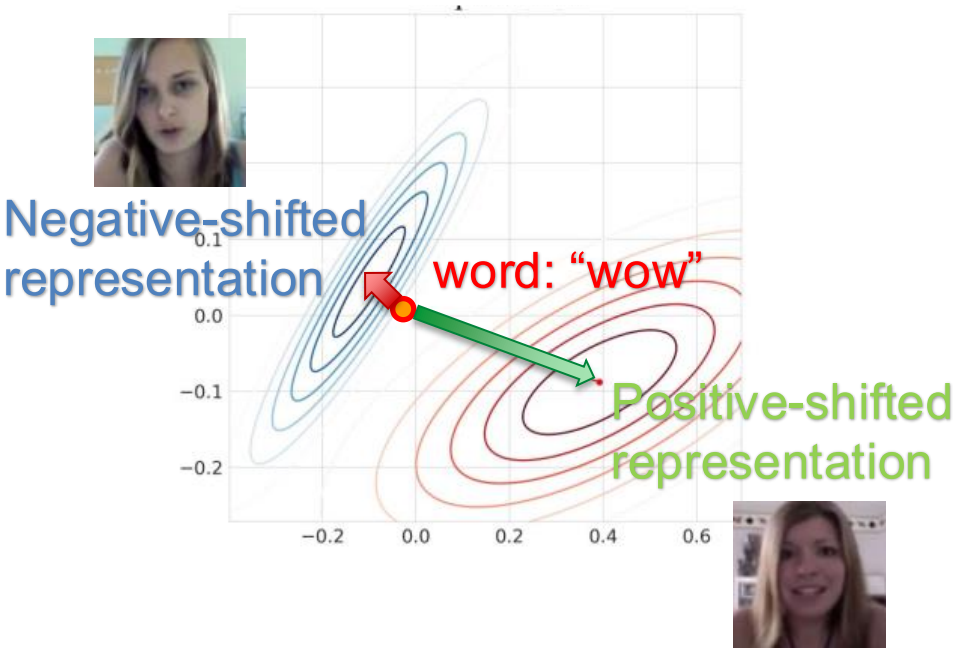
Modality-Shifting Fusion



Example with language modality:

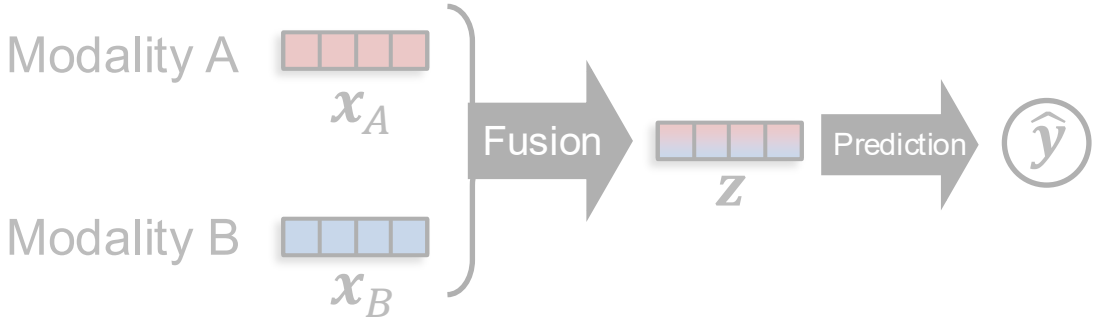
Primary modality: language

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]

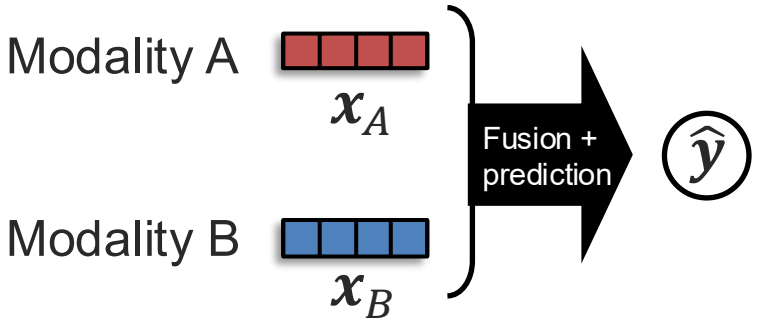
Nonlinear Fusion



Nonlinear fusion:

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

where f could be a neural network or any nonlinear model



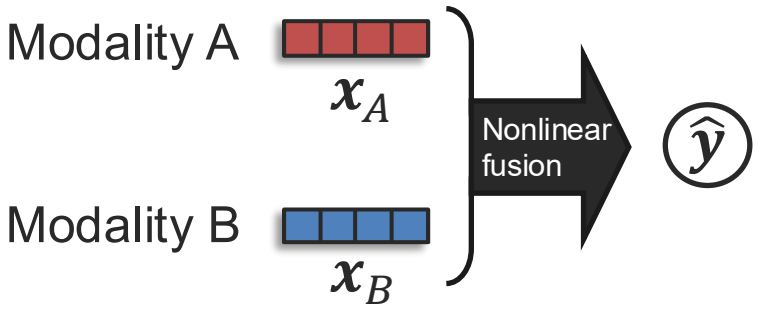
➡ This could be seen as *early fusion*:

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

… but will our neural network learn the nonlinear interactions?

multisensory intelligence

Measuring Non-Additive Interactions



Nonlinear fusion:

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

Projection?

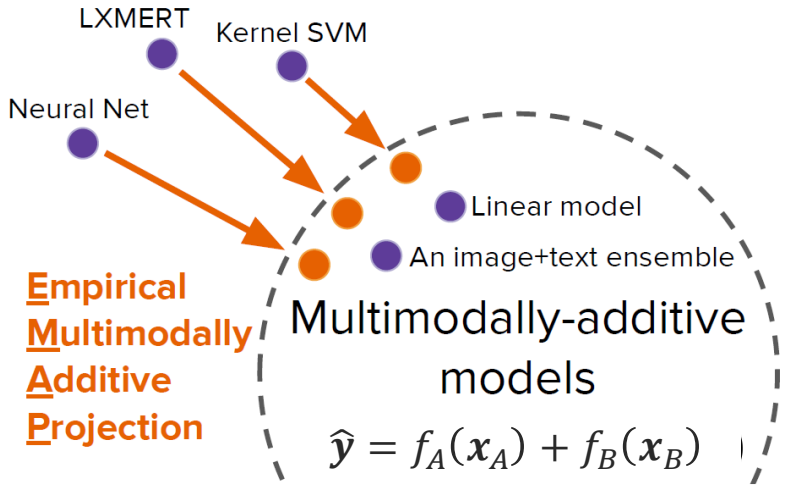
Additive fusion:

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

Projection from nonlinear to additive (using EMAP):

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

Modality A Modality B

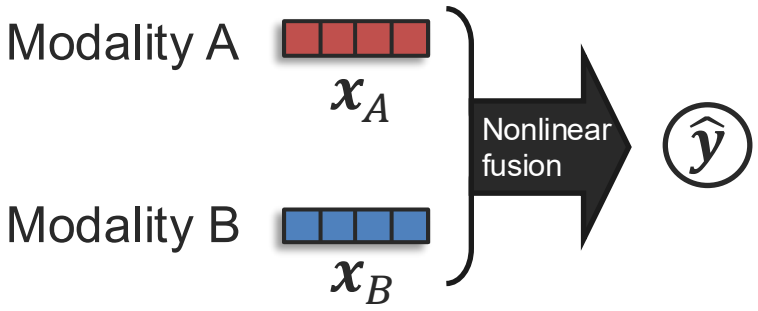


Additive fusion (approximation)

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

multisensory intelligence

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!

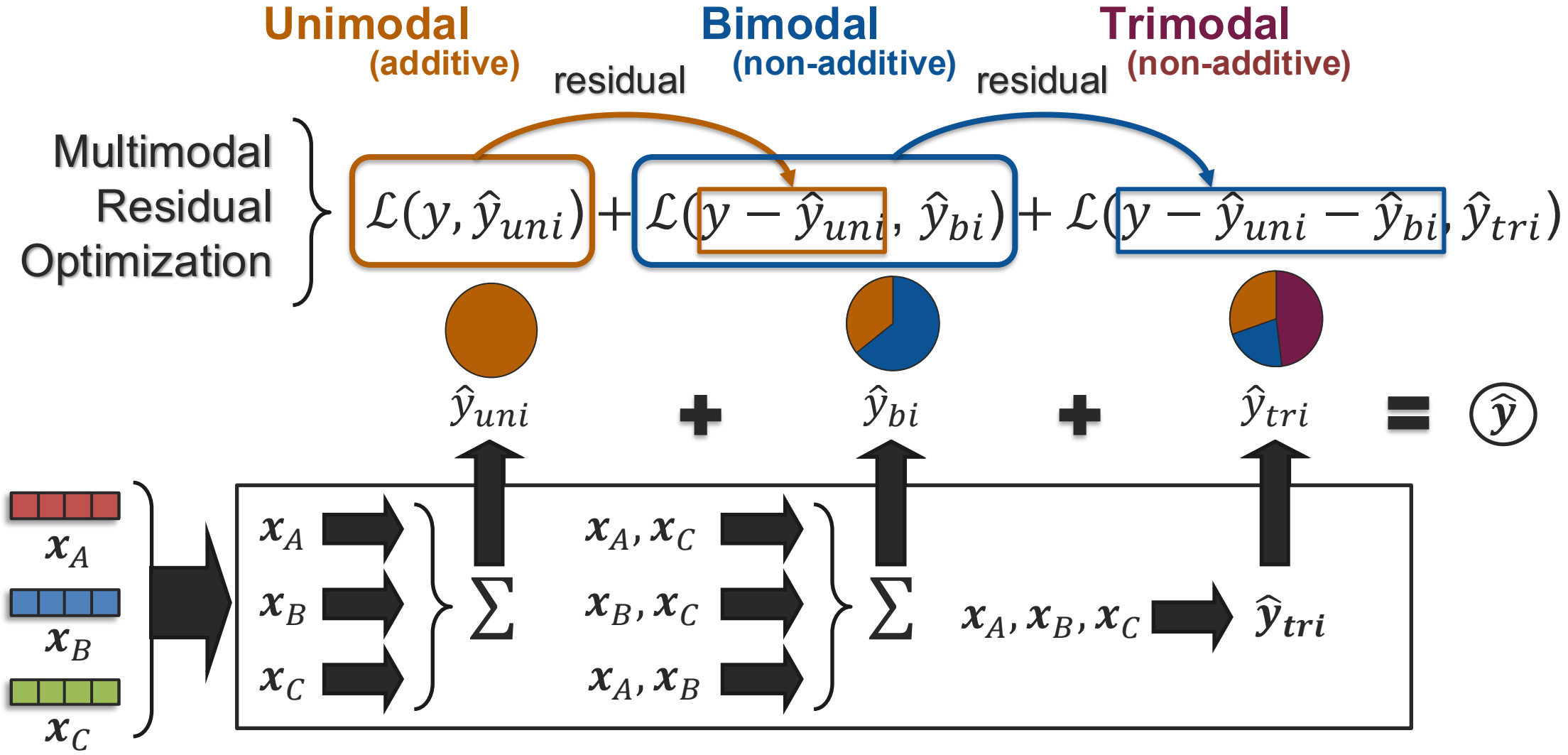
Differences are small!!!

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

multisensory intelligence

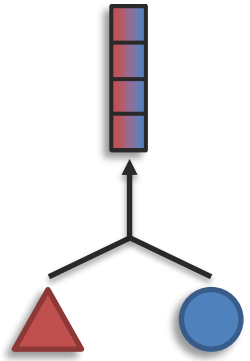
Non-Additive Interactions

Idea: prioritize simpler interactions



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

Summary: How To Multimodal Fusion



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

Homogenous modalities

Late fusion
Additive fusion
Multiplicative fusion
Tensor fusion
Polynomial fusion
Gated fusion
Modality-shift fusion
Nonlinear fusion
Very early fusion
Dynamic early fusion
Heterogeneity-aware
Improving optimization
Improving robustness

Heterogenous modalities

multisensory intelligence

Assignments for This Coming Week

Project proposal due today. Meet with me at 4-5pm if you want feedback.

I want to meet every group at least once regarding their project ideas either today or this Thursday.

HW2 released last week, due next Wednesday 3/4.