How to AI (Almost) Anything

Lecture 5 – 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

For project:

- I gave feedback and assigned primary TA.
- Meet with me and primary TA every other week.
- Should have finalized main ideas and experimental setup, have baseline models working, progress towards implementing new ideas.

Reading assignment due tomorrow Wednesday (3/12).

This Thursday (3/13): third reading discussion on **multimodal alignment**.

What views for contrastive learning

Platonic representation hypothesis

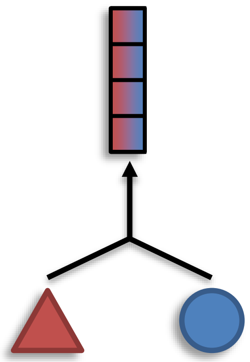
multisensory intelligence

Today's lecture

1. Basics of multimodal fusion
2. Early, intermediate, late fusion
3. Multiplicative and dynamic fusion
4. Complex fusion and improving optimization

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

Sub-Challenge 1a: Representation Fusion



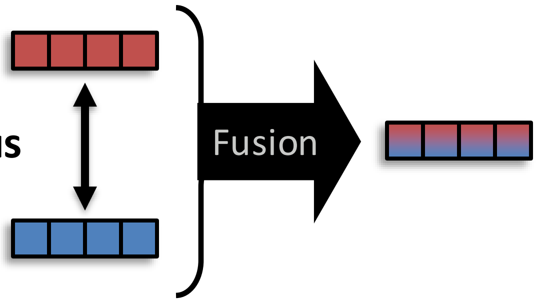
Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities.

Fusion with abstract modalities:

Modality A

Homogeneous

Modality B

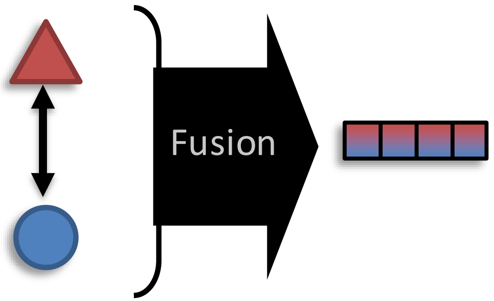


Fusion with raw modalities:

Modality A

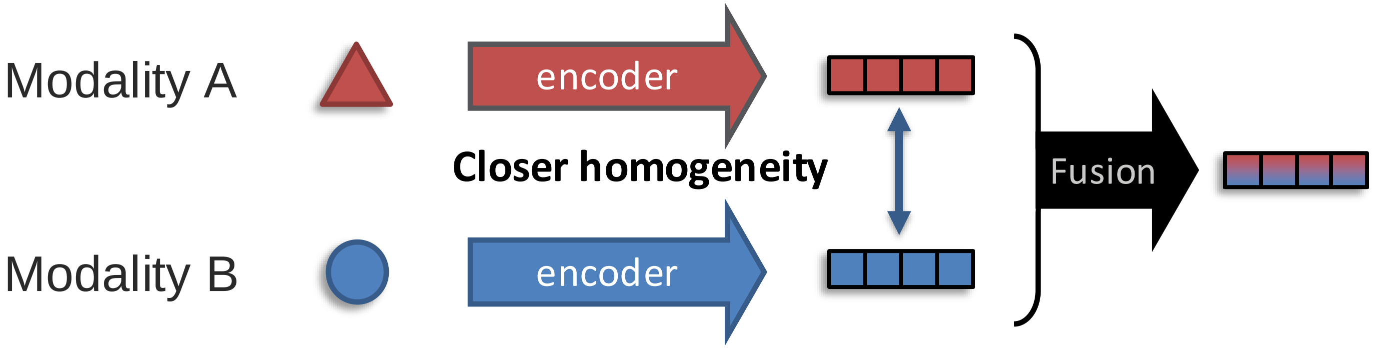
Heterogeneous

Modality B

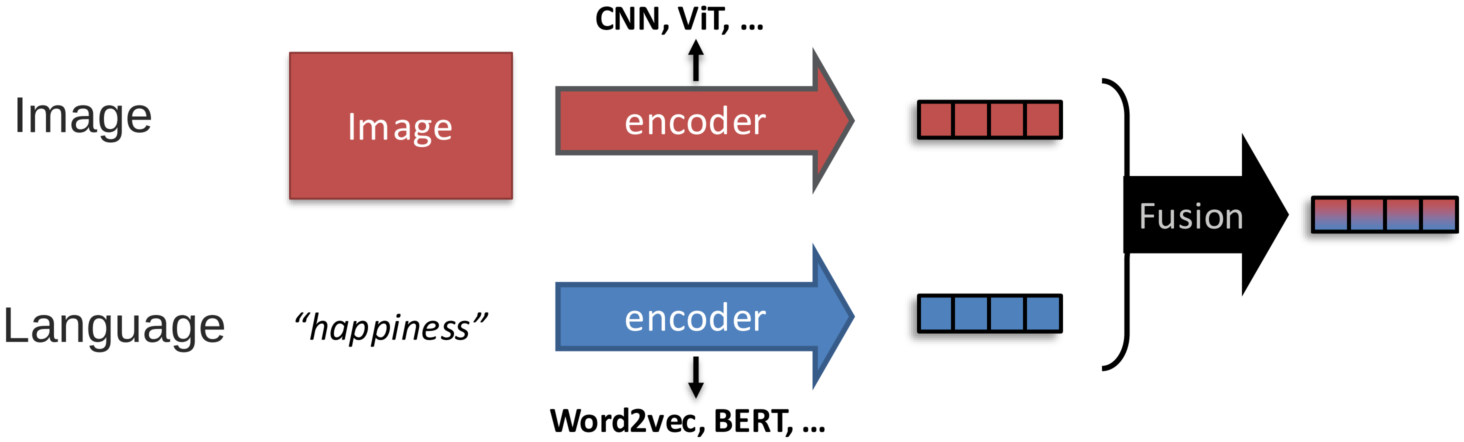


[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Fusion with Abstract Modalities



Example:



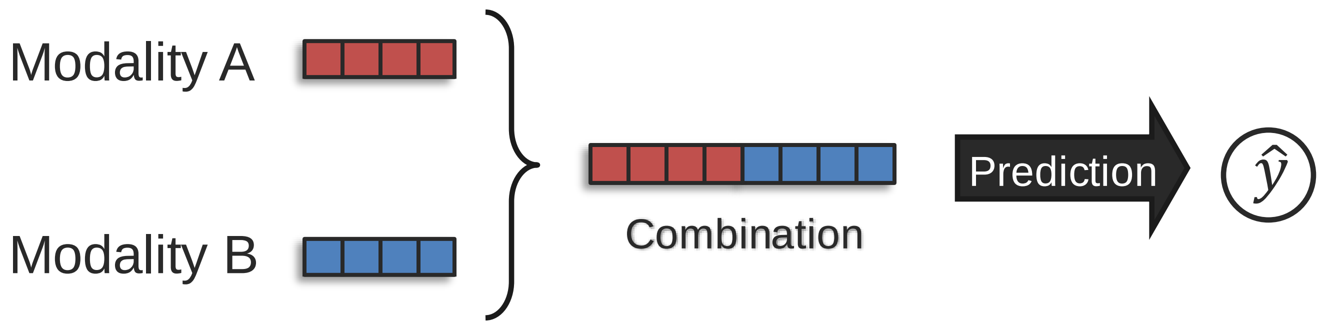
➡ Unimodal encoders can be jointly learned with fusion network, or pre-trained

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

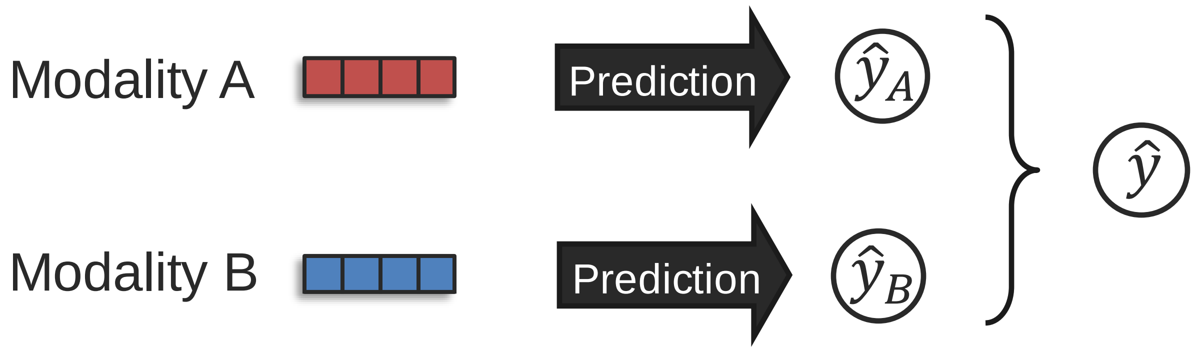
multisensory intelligence

Early and Late Fusion

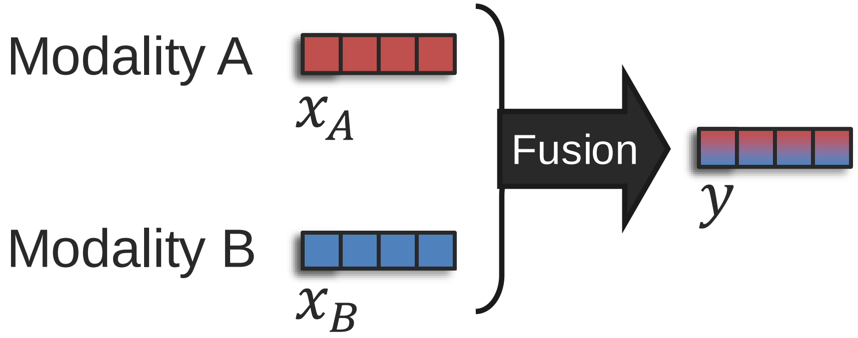
Early fusion:



Late fusion:



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:

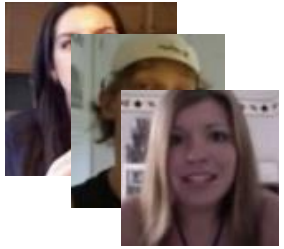
$$y = \underbrace{w_0}_{\substack{\text{intercept}\\\text{(bias term)}}} + \underbrace{w_1 x_A + w_2 x_B}_{\substack{\text{Additive}\\\text{terms}}} + \underbrace{w_3(x_A \times x_b)}_{\substack{\text{Multiplicative}\\\text{term}}} + \underbrace{\epsilon}_{\substack{\text{error}\\\text{(residual term)}}}$$

[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)

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

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

Linear regression:

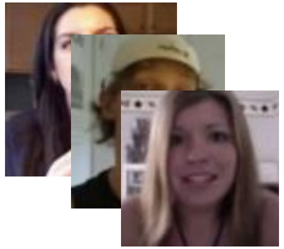
$$y = \underset{\substack{\downarrow \\ \text{intercept} \\ \text{(bias term)}}}{w_0} + \underbrace{w_1 x_A + w_2 x_B}_{\text{Additive terms}} + \underbrace{w_3(x_A \times x_b)}_{\text{Multiplicative term}} + \underset{\substack{\downarrow \\ \text{error} \\ \text{(residual term)}}}{\epsilon}$$

- 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)

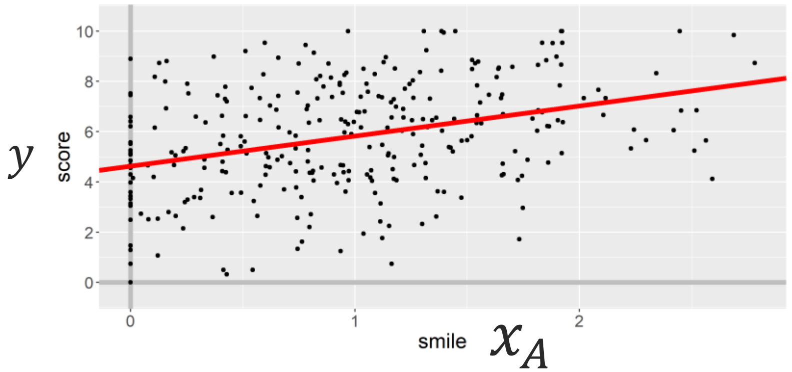
H1: *Does smiling reveal what the audience score was?*

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

Linear regression:

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

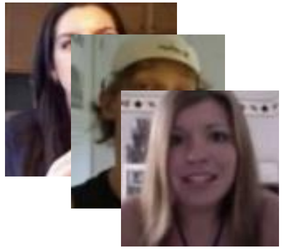
slope



	Estimate
w_0	4.63
w_1	1.20

Linear Fusion Case

300 book reviews



y: audience score

x_A: percentage of smiling

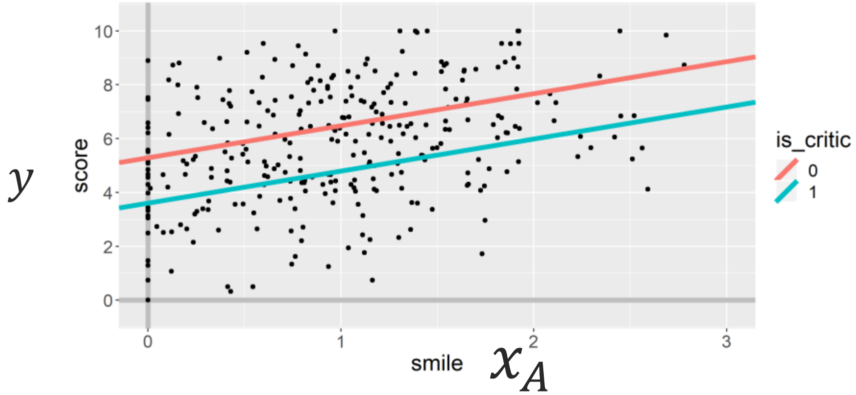
x_B: professional status (0=non-critic, 1=critic)

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

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

Linear regression:

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



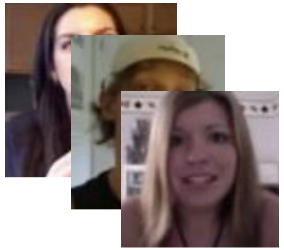
	Estimate
w_0	5.29
w_1	1.19
w_2	-1.69

Positive effect

Negative effect

Linear Fusion Case

300 book reviews



y: audience score

x_A: percentage of smiling

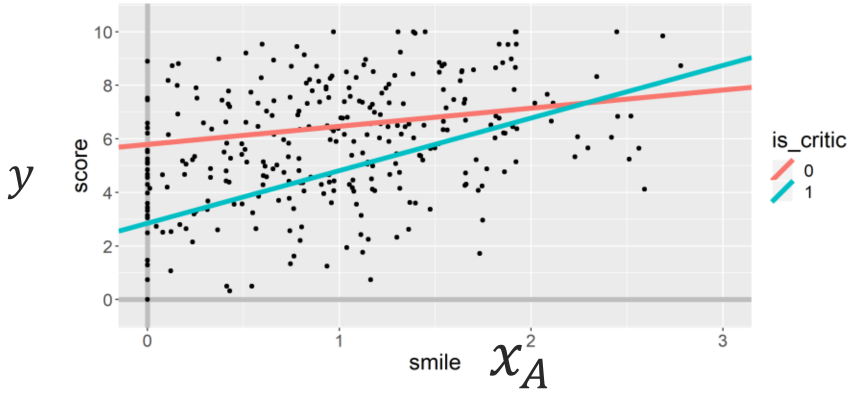
x_B: professional status
(0=non-critic, 1=critic)

H1: Does smiling reveal what the audience score was?

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

Linear regression:

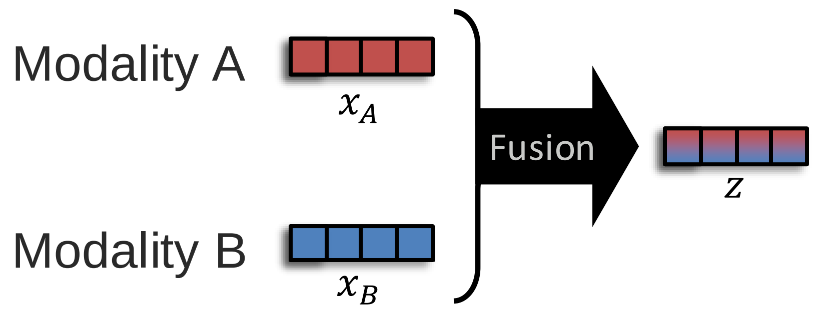
$$y = w_0 + \boxed{w_1} x_A + \boxed{w_2} x_B + \boxed{w_3} (x_A \times x_b) + \epsilon$$



	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:

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

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:

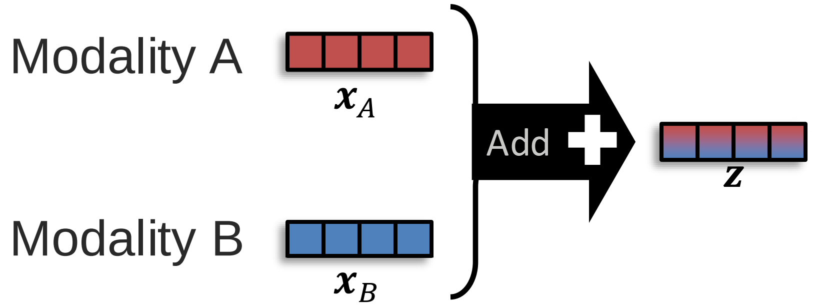
$$z = w_1 x_A + w_2 x_B + w_3(x_A \times x_b) + \epsilon$$

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

Additive Fusion

Back to multivariate case!

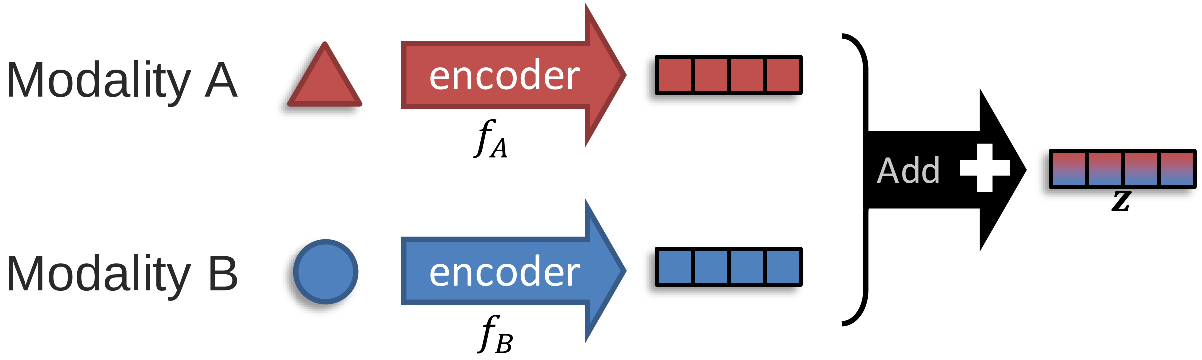
(multi-dimensional features)



Additive fusion:

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

With unimodal encoders:



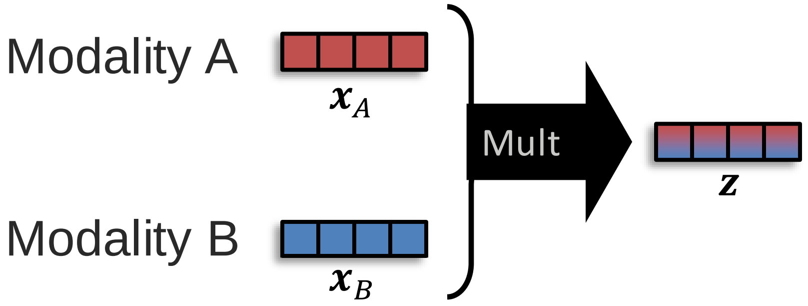
Additive fusion:

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

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

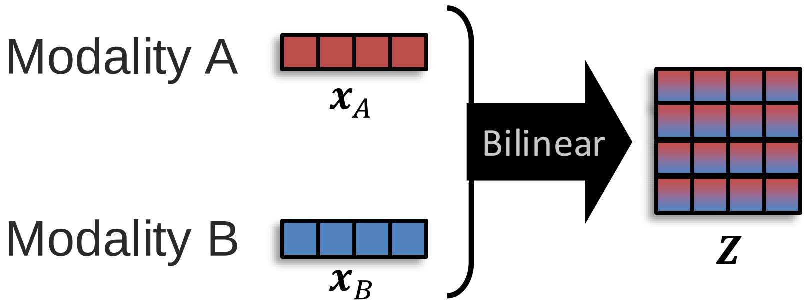
[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]

Multiplicative Fusion



Multiplicative fusion:

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



Bilinear Fusion:

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

Multimodal Transformers

It's just a privilege to watch your mind at work.

privilege
mind

vocal emphasis
eye-roll

vocal emphasis
eye-roll

Vision-to-language attention

	vocal emphasis			eye-roll
privilege	0.7			0.3
mind	1.0			

privilege
mind

New **language** representation interacting with vision

Sarcasm

3 x 4

(weighted) outer product

(row) normalize to 0-1

3 x d   d x 4   4 x d

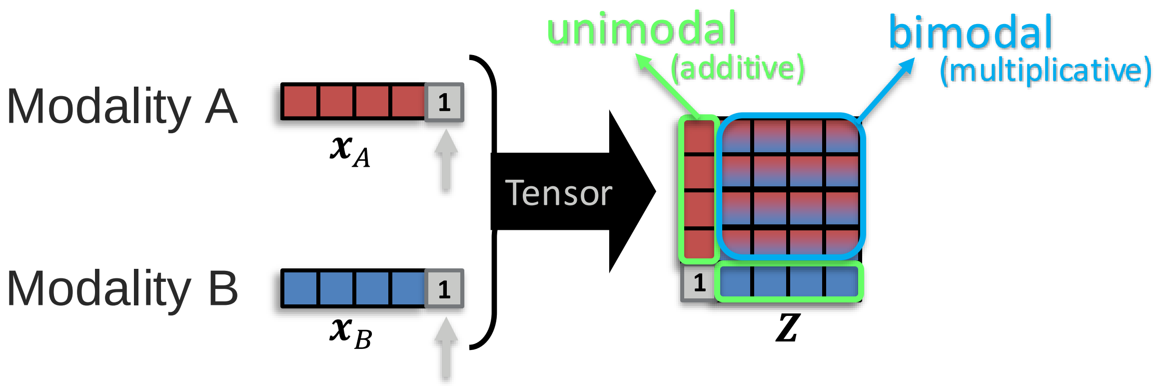
$$h = \text{softmax}\left(\frac{X_1 W_q W_k^T X_2^T}{\sqrt{d}}\right) X_2 W_v$$

normalize wrt dimension *d*

[Liang et al., Multimodal Language Analysis with Recurrent Multistage Fusion. EMNLP 2018]
[Tsai, Bai, Liang et al., Multimodal Transformer for Unaligned Multimodal Language Sequences. ACL 2019]

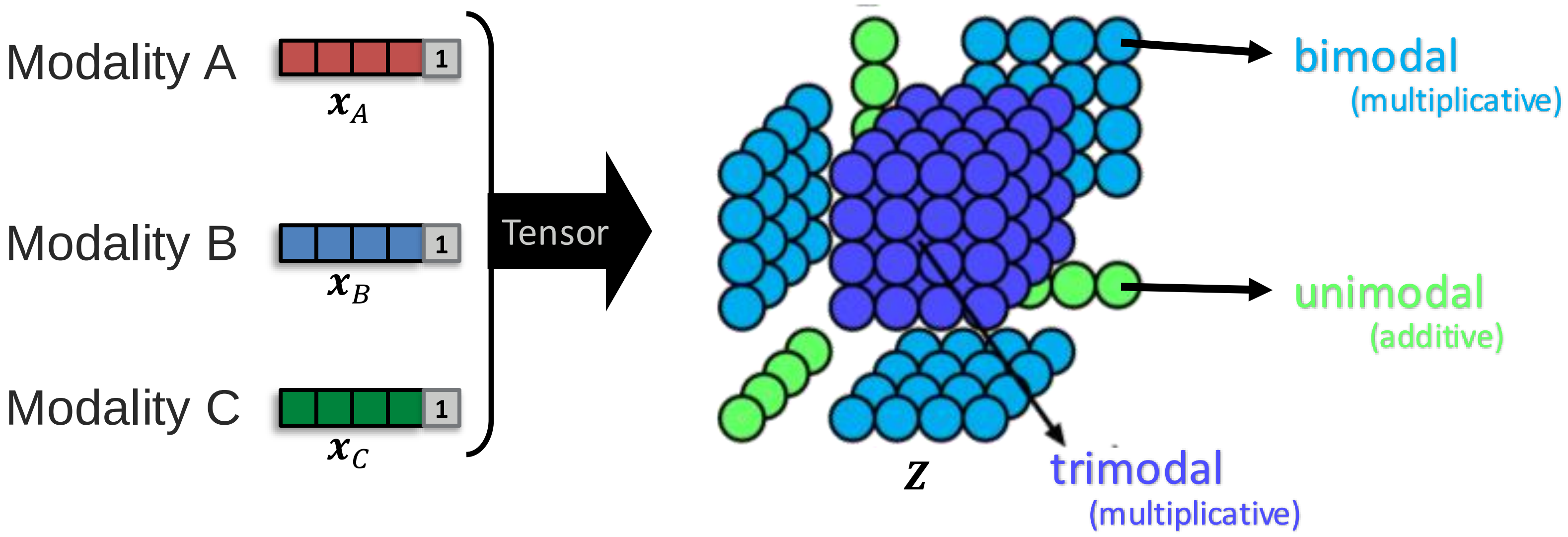
multisensory intelligence

Tensor Fusion



Tensor Fusion (bimodal):

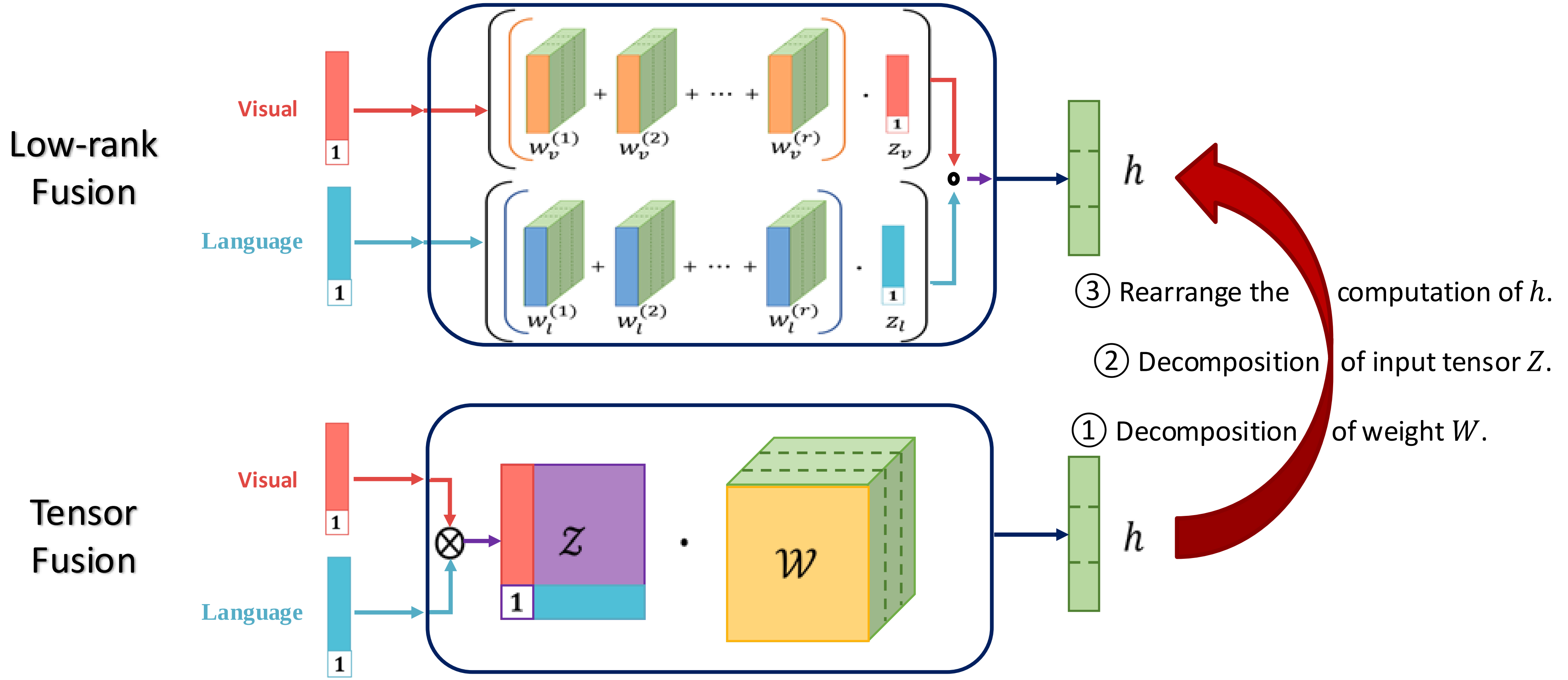
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

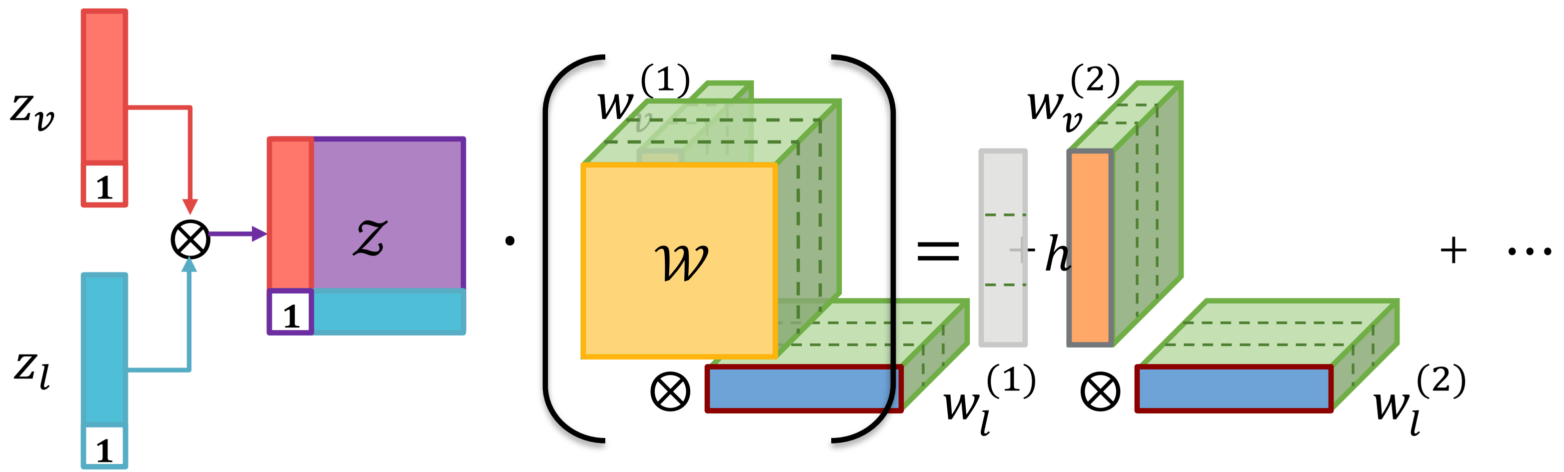
[Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis. EMNLP 2017]

Low-rank Fusion



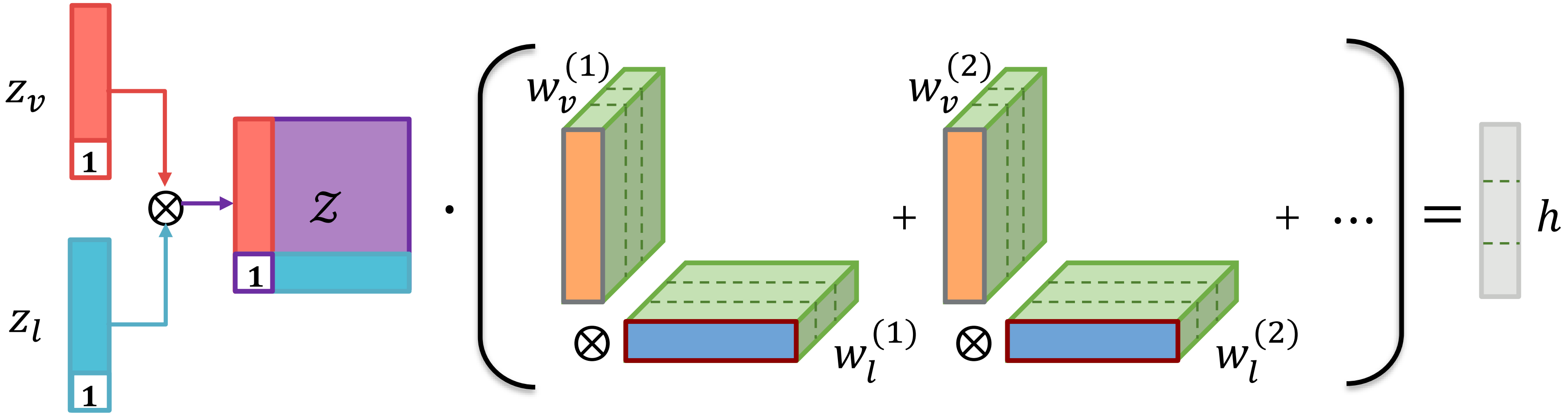
[Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors. ACL 2018]
[Hu et al., LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022]

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]

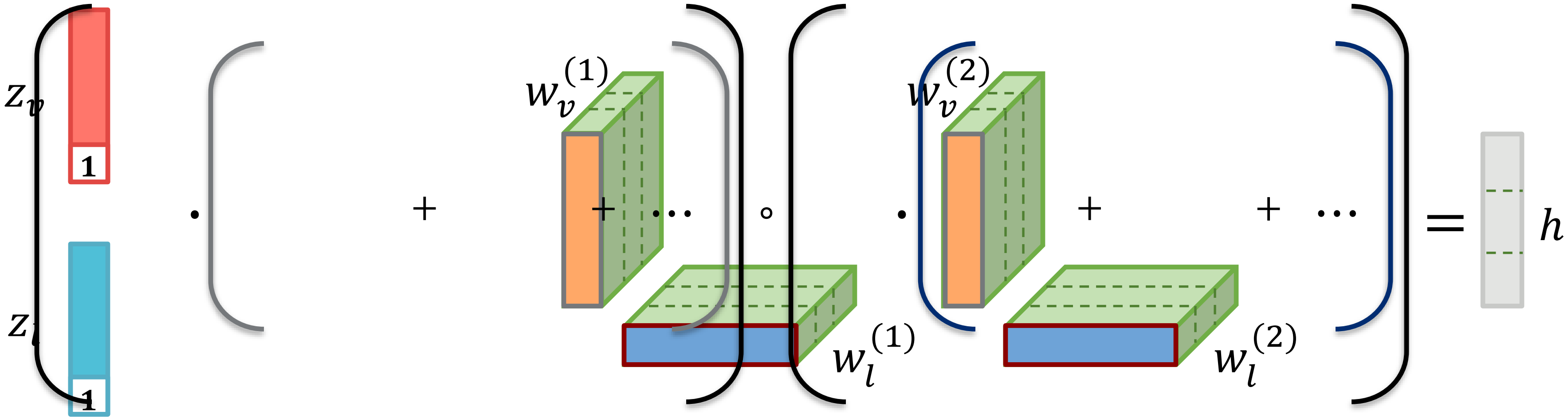
Low-rank Fusion



[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]

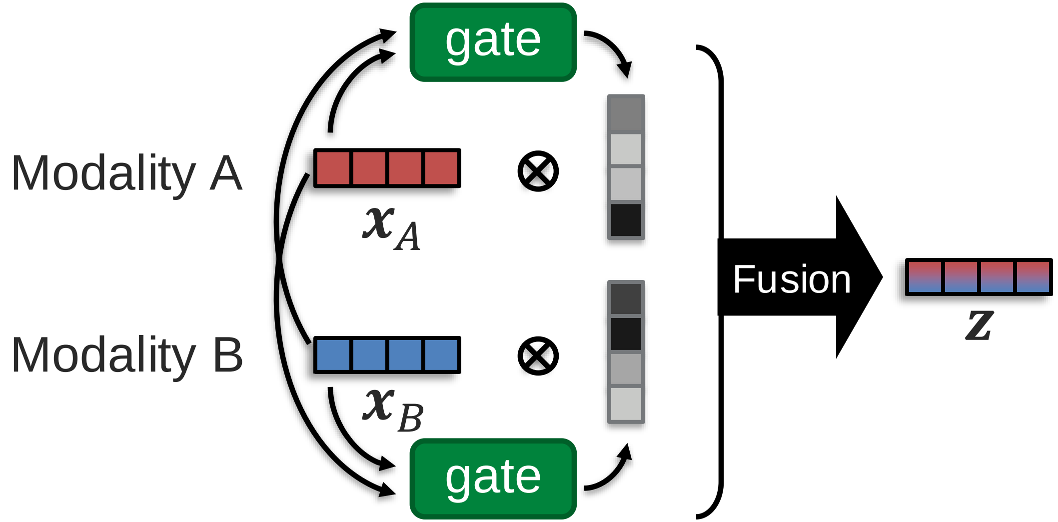
multisensory intelligence

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]

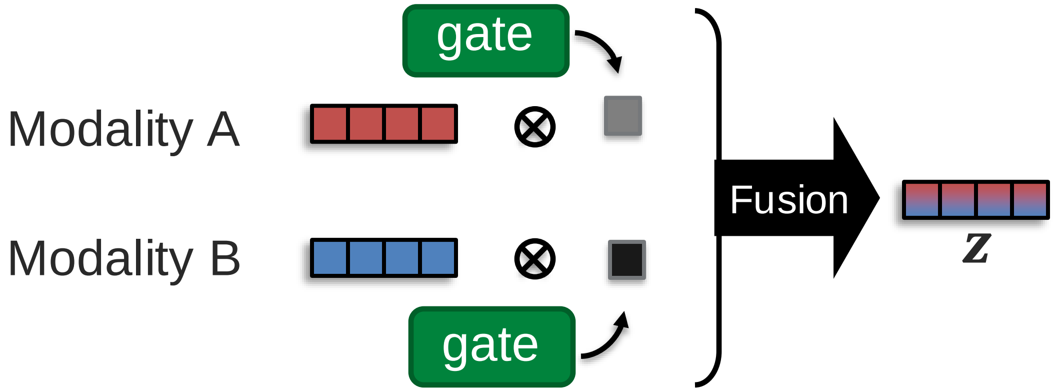
Gated Fusion



Example with additive fusion:

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

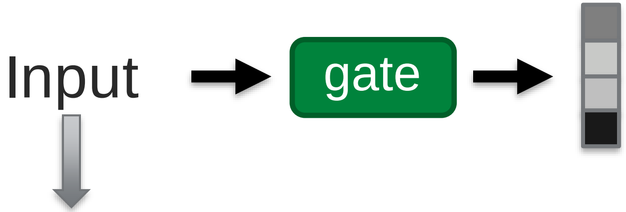
➡ g_A and g_B can be seen as attention functions



➡ Gating output can be one weight for the whole modality

[Arevalo et al., Gated Multimodal Units for Information Fusion. ICLR-workshop 2017]

Gated Fusion



What should it be?

Target modality

Other modality

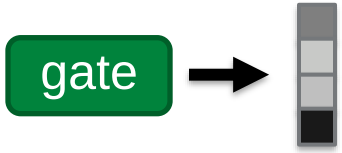
All modality

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

…or with a more positive view:

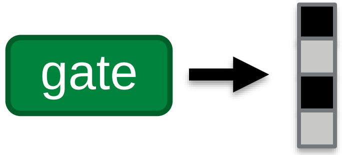
"Neural network designed to select preferable signal to move forward" (attention)

Soft attention



Easier to compute derivative (gradient)

Hard attention

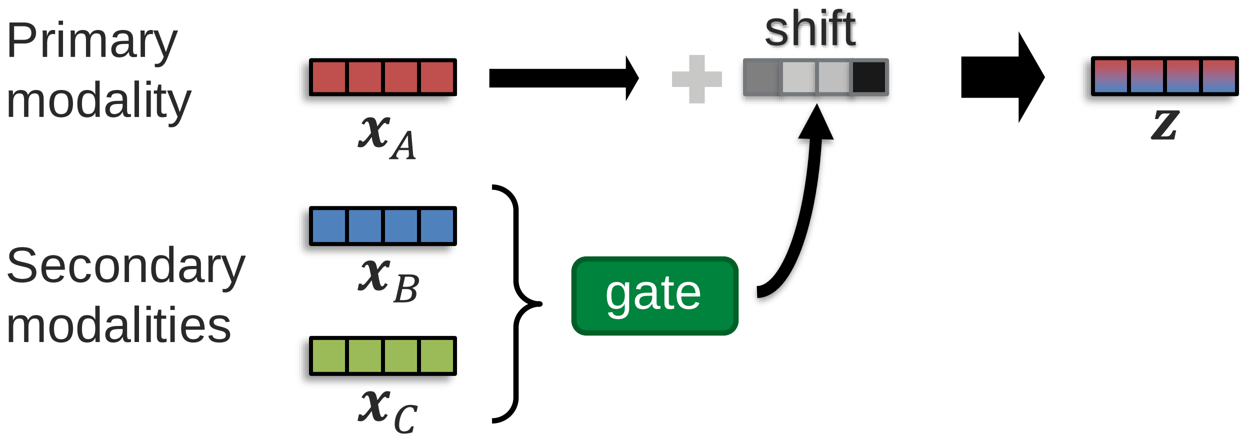


Derivative is harder (e.g., use reinforcement learning)

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

multisensory intelligence

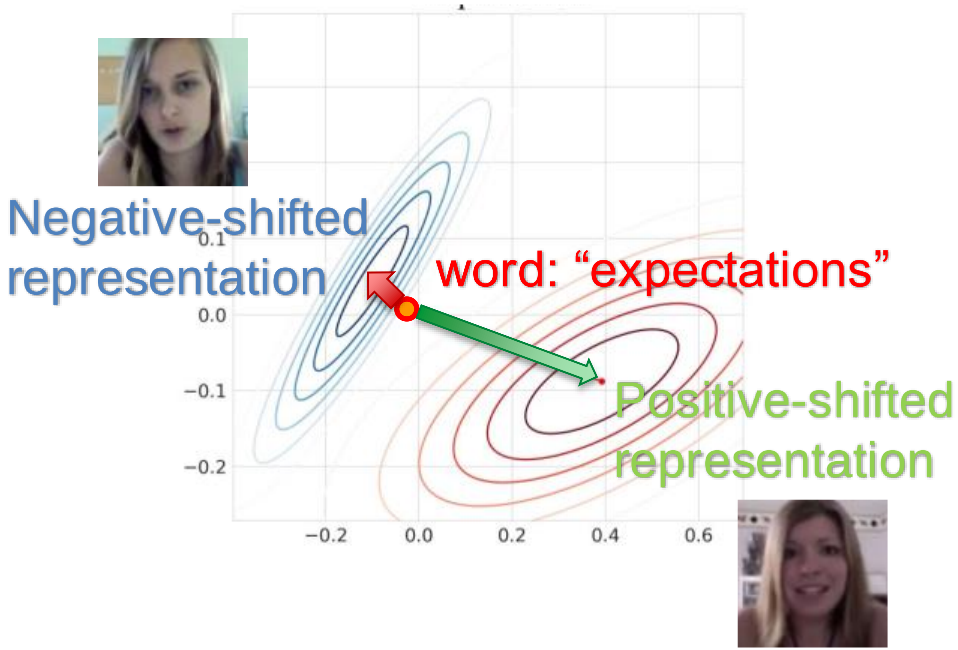
Modality-Shifting Fusion



Example with language modality:

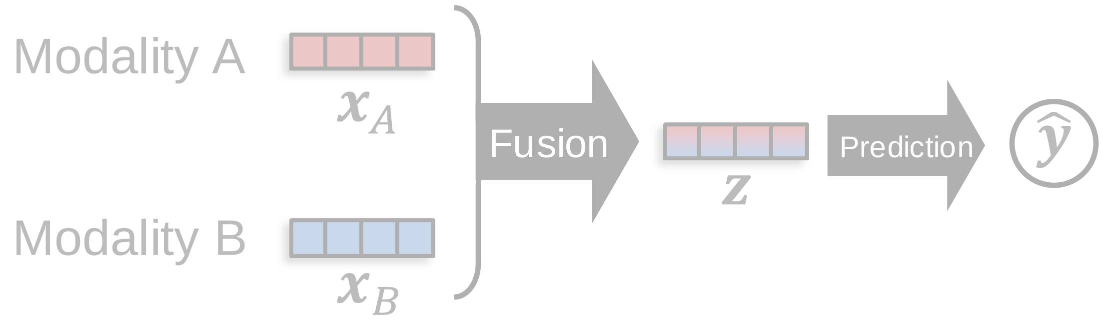
Primary modality: language

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]

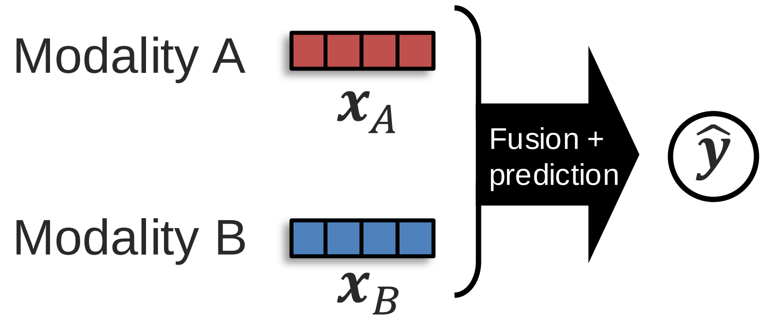
Nonlinear Fusion



Nonlinear fusion:

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

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

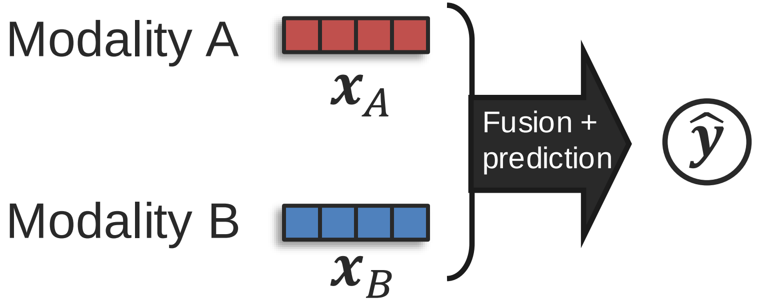


This could be seen as *early fusion*:

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

… but will our neural network learn the nonlinear interactions?

Measuring Non-Additive Interactions



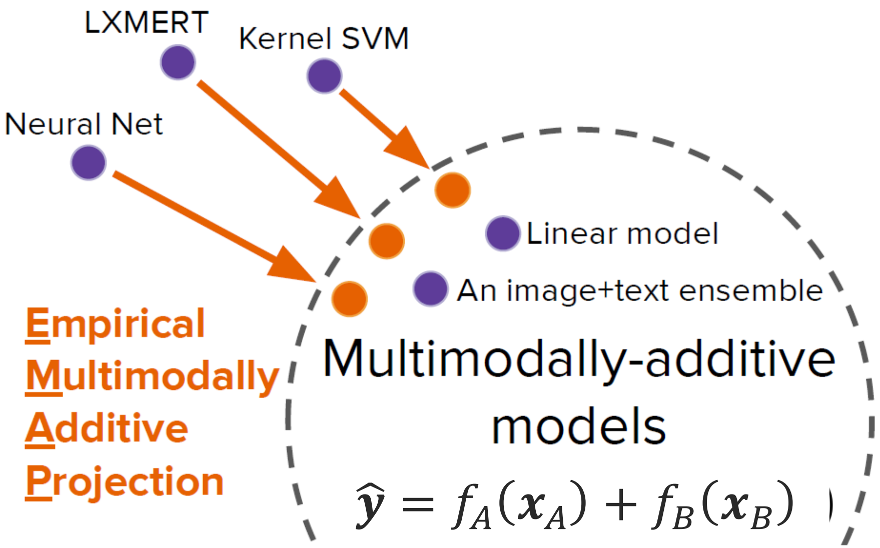
Nonlinear fusion:

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

Projection?

Additive fusion:

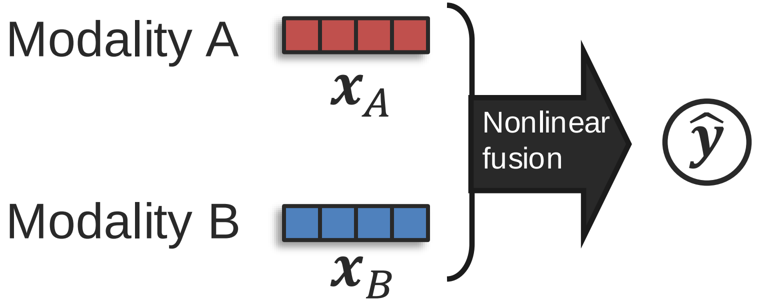
$$\hat{y} = f_A(\boldsymbol{x}_A) + f_B(\boldsymbol{x}_B)$$



Empirical
Multimodally
Additive
Projection

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

Measuring Non-Additive Interactions



Nonlinear fusion:

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

Projection?

Additive fusion:

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

Projection from nonlinear to additive (using EMAP):

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

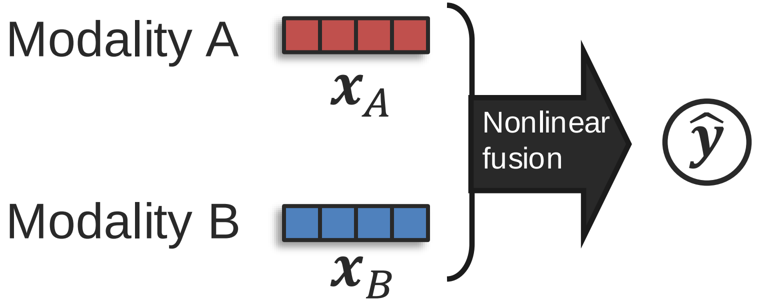
Modality A + Modality B

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

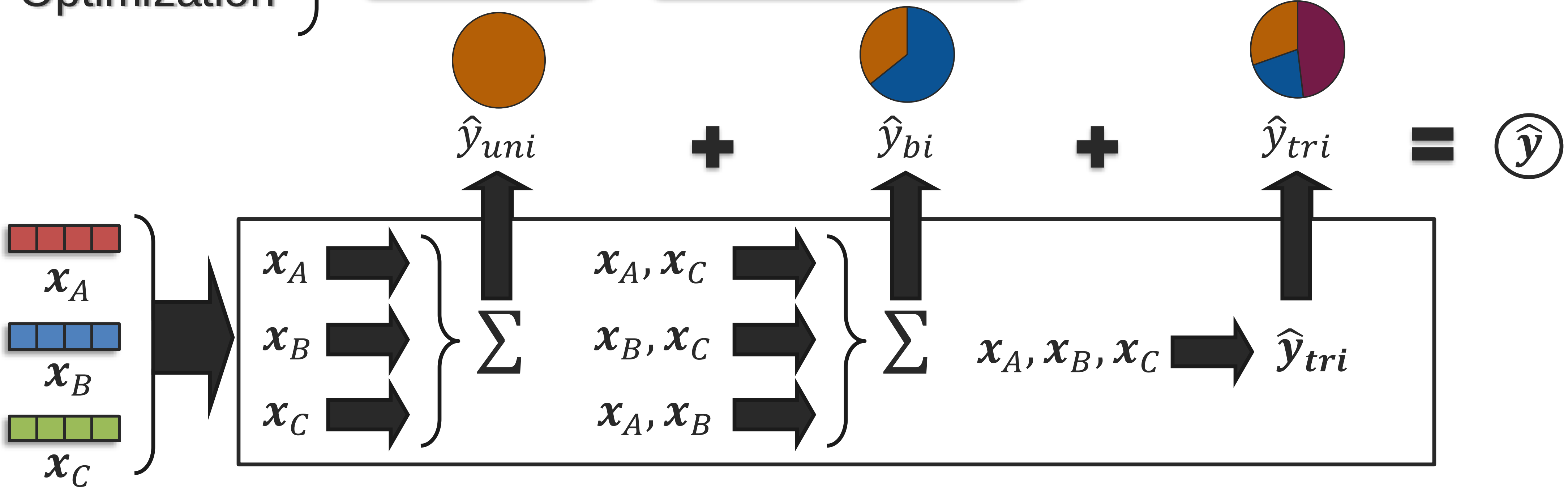
Non-Additive Interactions

Idea: prioritize simpler interactions

Unimodal (additive) residual **Bimodal** (non-additive) residual **Trimodal** (non-additive)

Multimodal Residual Optimization

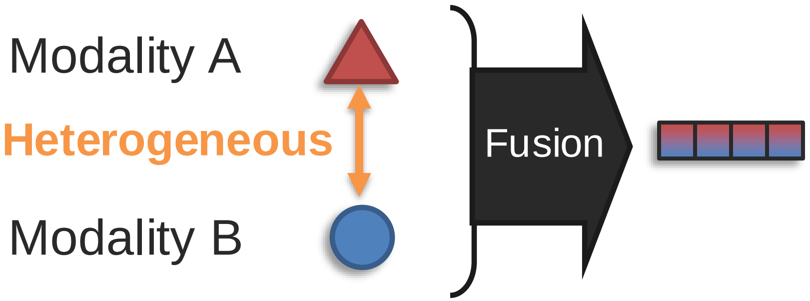
$$\mathcal{L}(y, \hat{y}_{uni}) + \mathcal{L}(y - \hat{y}_{uni}, \hat{y}_{bi}) + \mathcal{L}(y - \hat{y}_{uni} - \hat{y}_{bi}, \hat{y}_{tri})$$



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

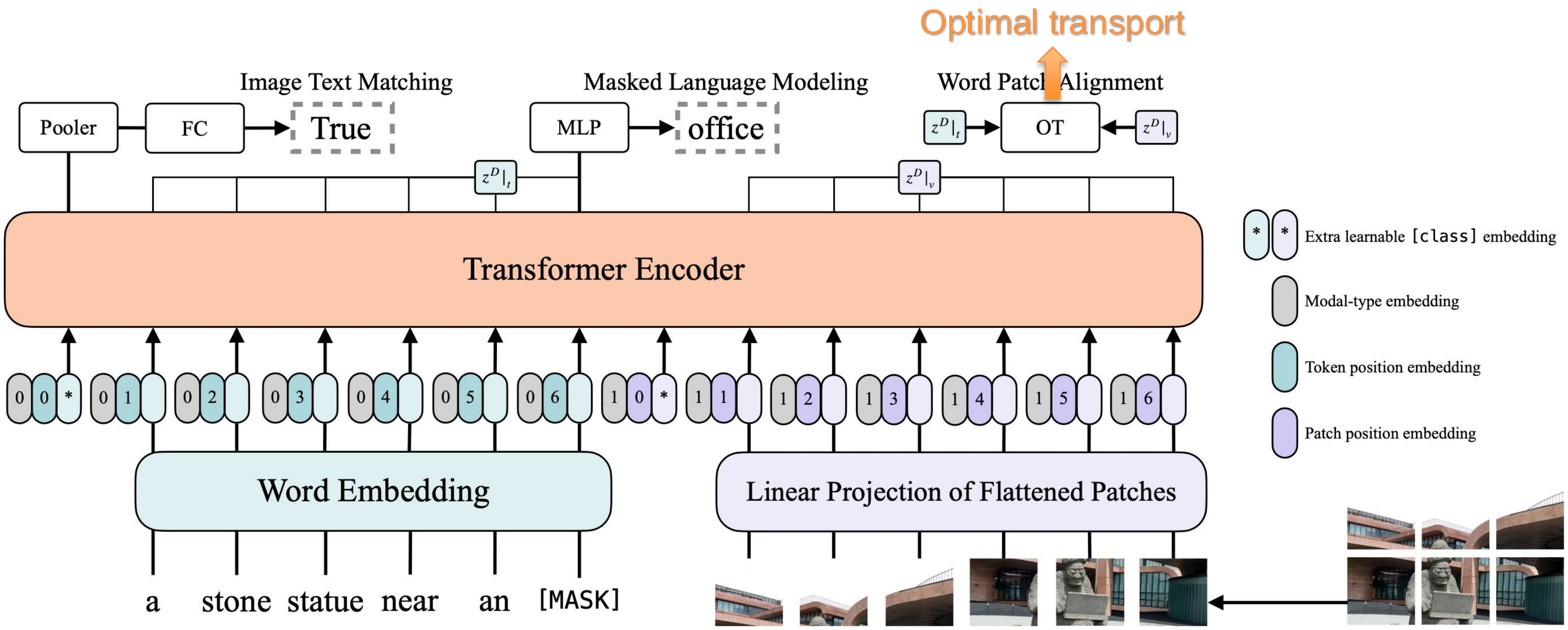
Fusion with Heterogeneous Modalities

Example: From feature fusion to early fusion



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

Visual-and-Language Transformer (ViLT) (≈ BERT + ViT)



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

[Kim et al., ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML 2021]

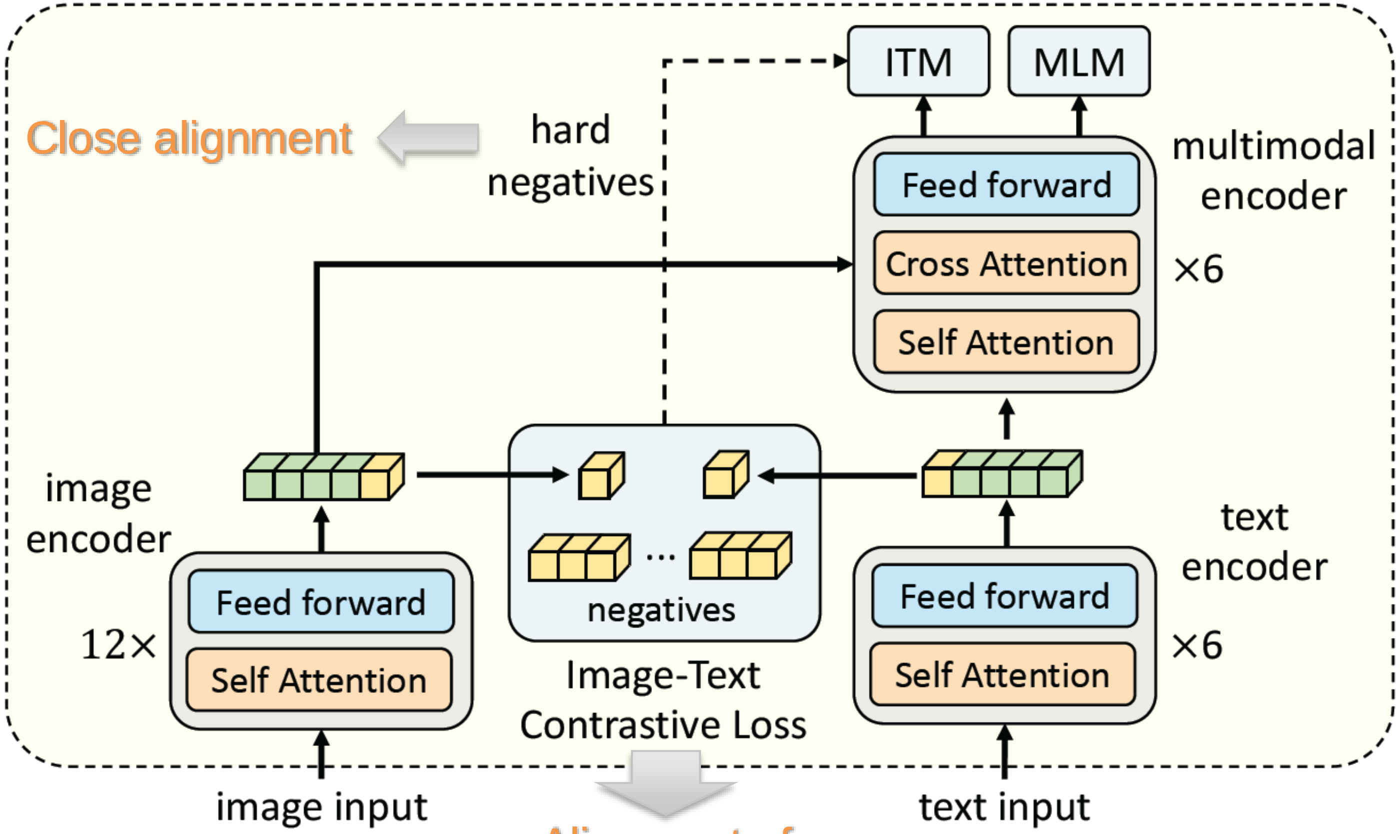
multisensory intelligence

Visual-and-Language Transformer (ViLT)

Example of alignment between modalities:

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]
[Kim et al., ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. ICML 2021]

ALBEF: Align Before Fusion (≈ BERT + ViT + CLIP-ish)



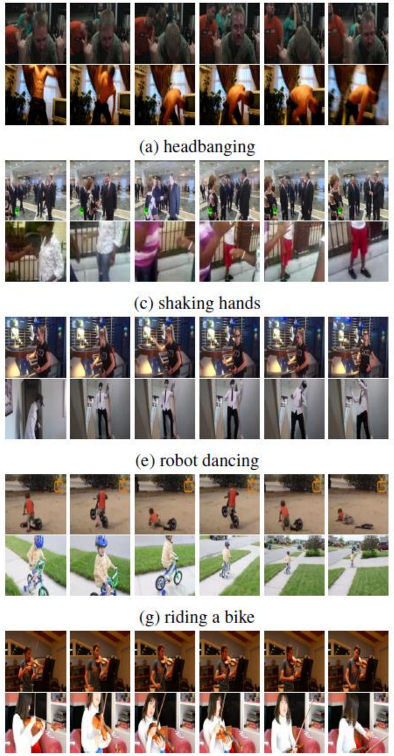
Alignment of segment-level embeddings → Related to mutual information

[original slide co-developed with Louis-Philippe Morency for CMU course 11-777]
[Li et al., Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. NeurIPS 2021]

multisensory intelligence

Nonlinear Fusion

Kinetics dataset



Adding more modalities should always help?

Modalities: **RGB** (video clips)
A (Audio features)
OF (optical flow - motion)

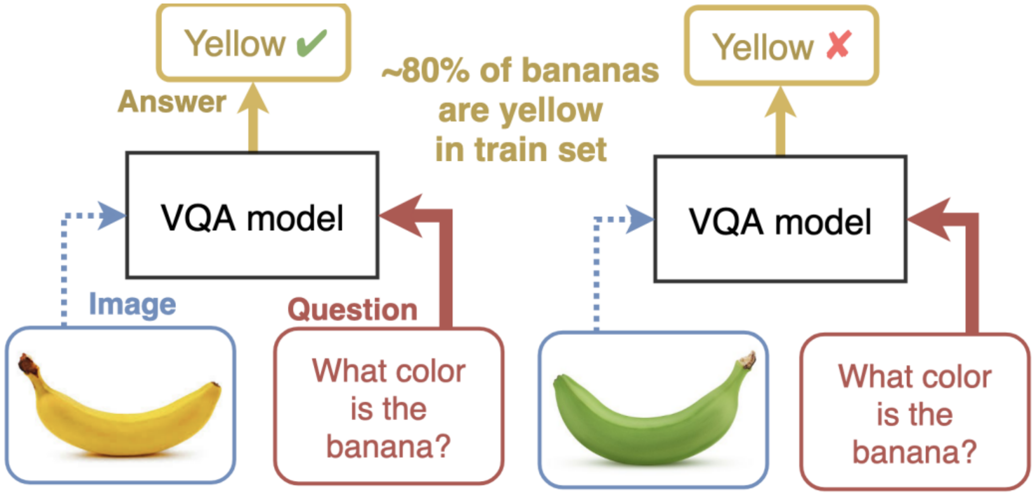
Dataset	Multi-modal	V@1	Best Uni	V@1	Drop
Kinetics	A + RGB	71.4	RGB	**72.6**	-1.2
	RGB + OF	71.3	RGB	**72.6**	-1.3
	A + OF	58.3	OF	**62.1**	-3.8
	A + RGB + OF	70.0	RGB	**72.6**	-2.6

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

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

Unimodal Biases

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

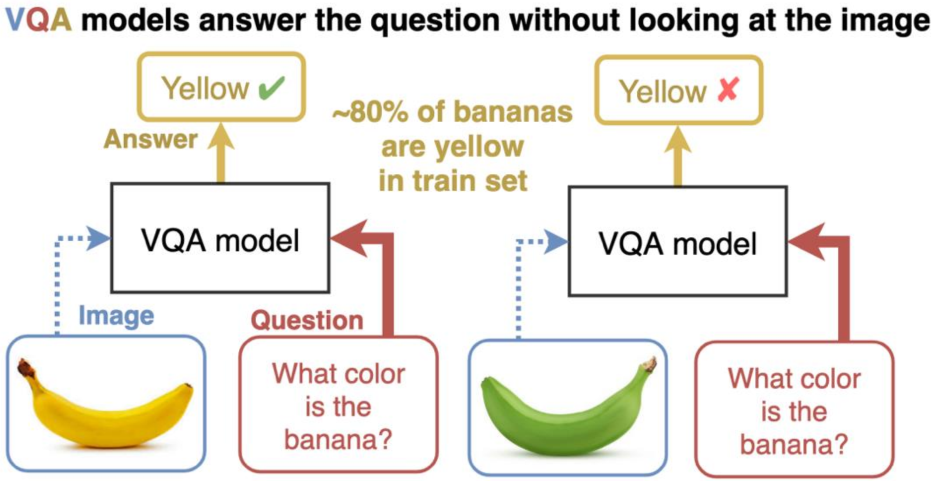


Finding: Image captioning models capture spurious correlations between gender and generated actions.

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

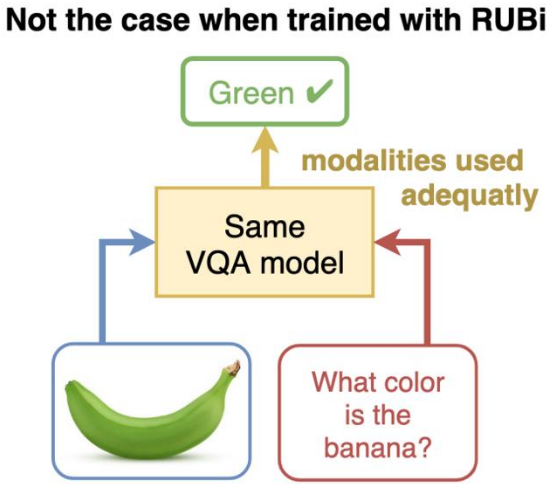
multisensory intelligence

Unimodal Biases



Balancing modalities

Balancing training



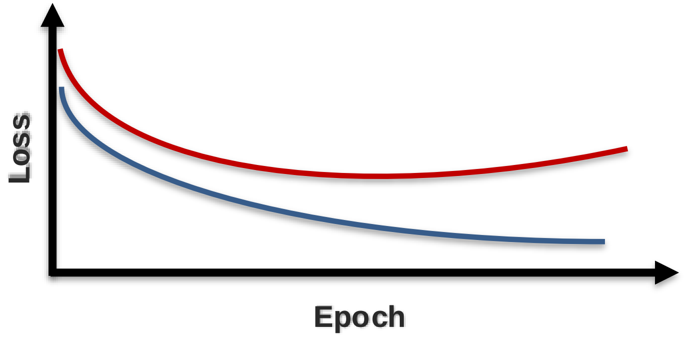
[Javaloy et al., Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. ICML 2022]
[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017]

multisensory intelligence

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)

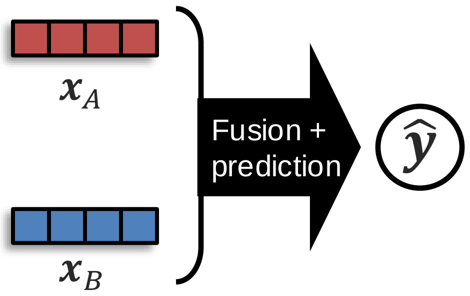
Gap between training and valid loss

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

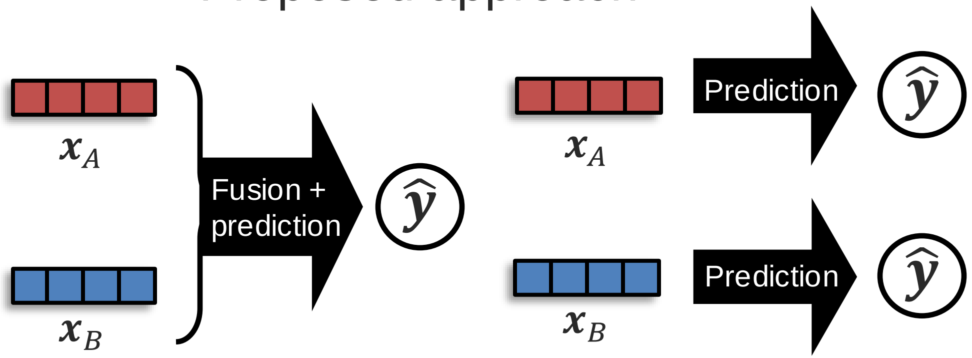
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]

Optimization Challenges

Conventional approach



Proposed approach



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

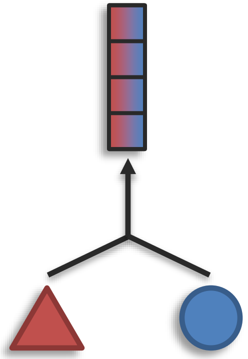
Reweight multimodal loss using unimodal OGR values

Allows to better balance generalization & overfitting rate of different modalities

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

multisensory intelligence

Summary: How To Multimodal Fusion



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

multisensory intelligence

Lecture Summary

1. Basics of multimodal fusion
2. Early, intermediate, late fusion
3. Multiplicative and dynamic fusion
4. Complex fusion and improving optimization

Assignments for This Coming Week

For project:

- I gave feedback and assigned primary TA.
- Meet with me and primary TA every other week.
- Should have finalized main ideas and experimental setup, have baseline models working, progress towards implementing new ideas.

Reading assignment due tomorrow Wednesday (3/12).

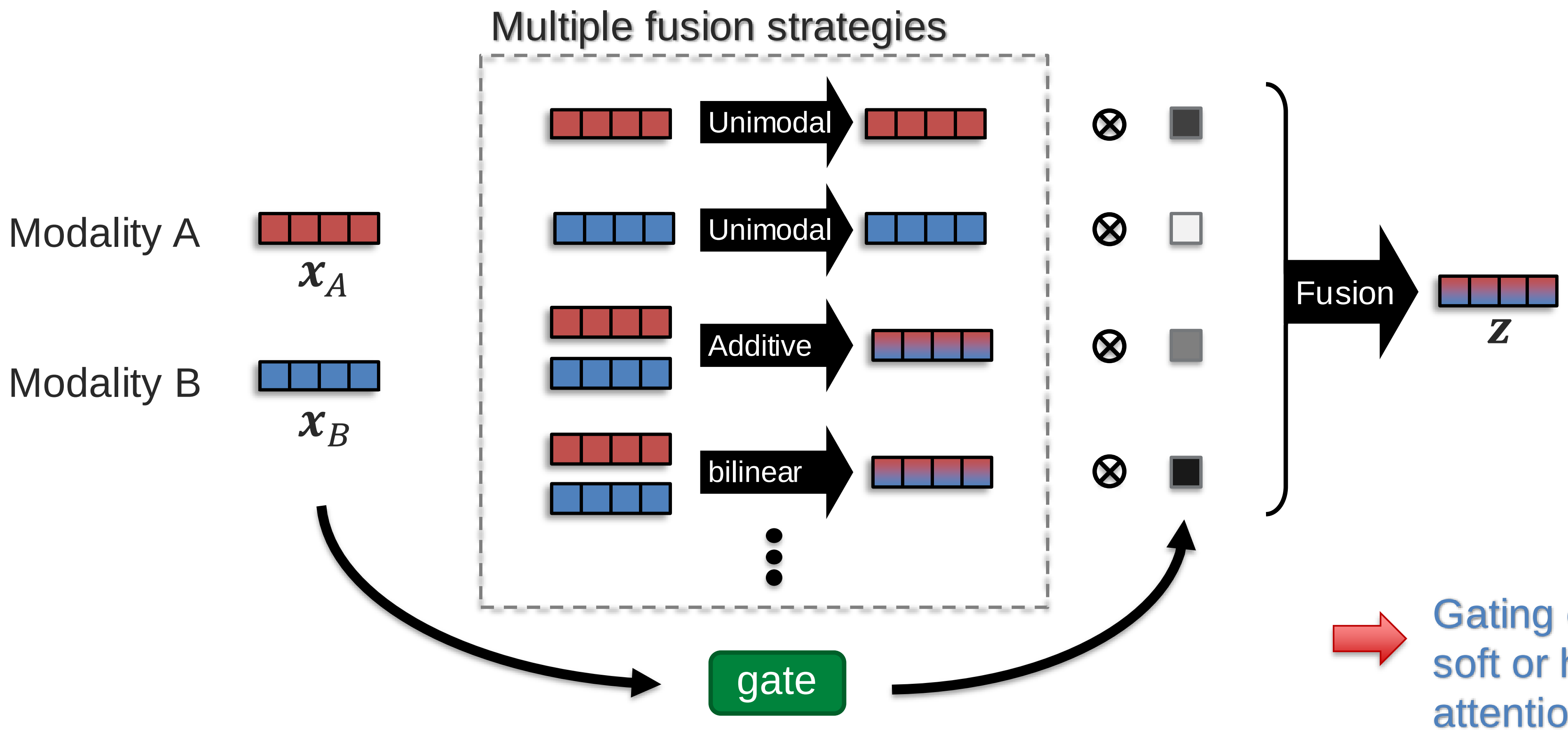
This Thursday (3/13): third reading discussion on **multimodal alignment**.

What views for contrastive learning

Platonic representation hypothesis

multisensory intelligence

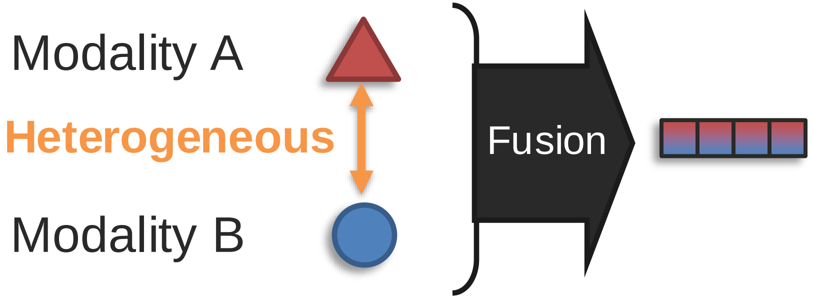
Dynamic Fusion



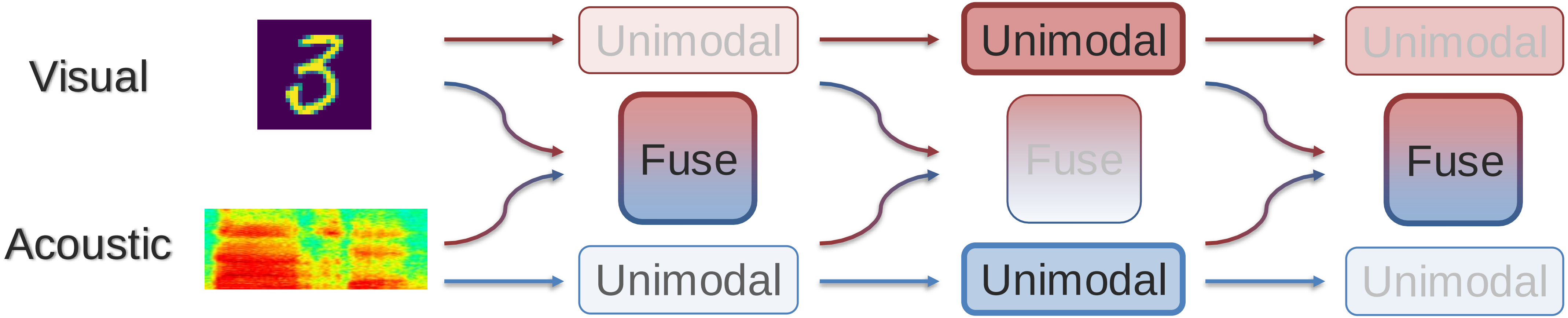
Gating can we with soft or hard attention

[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]

Dynamic Early Fusion



Idea: Deciding when to fuse in early fusion

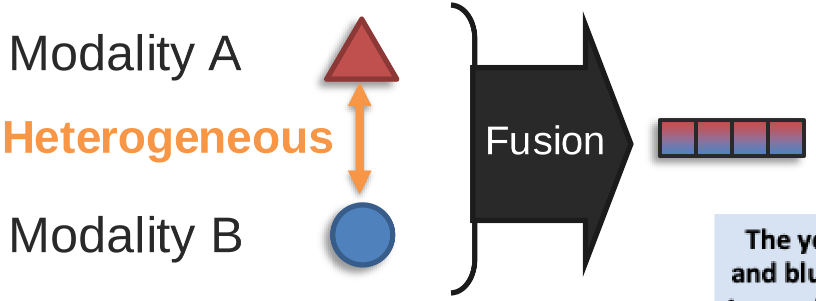


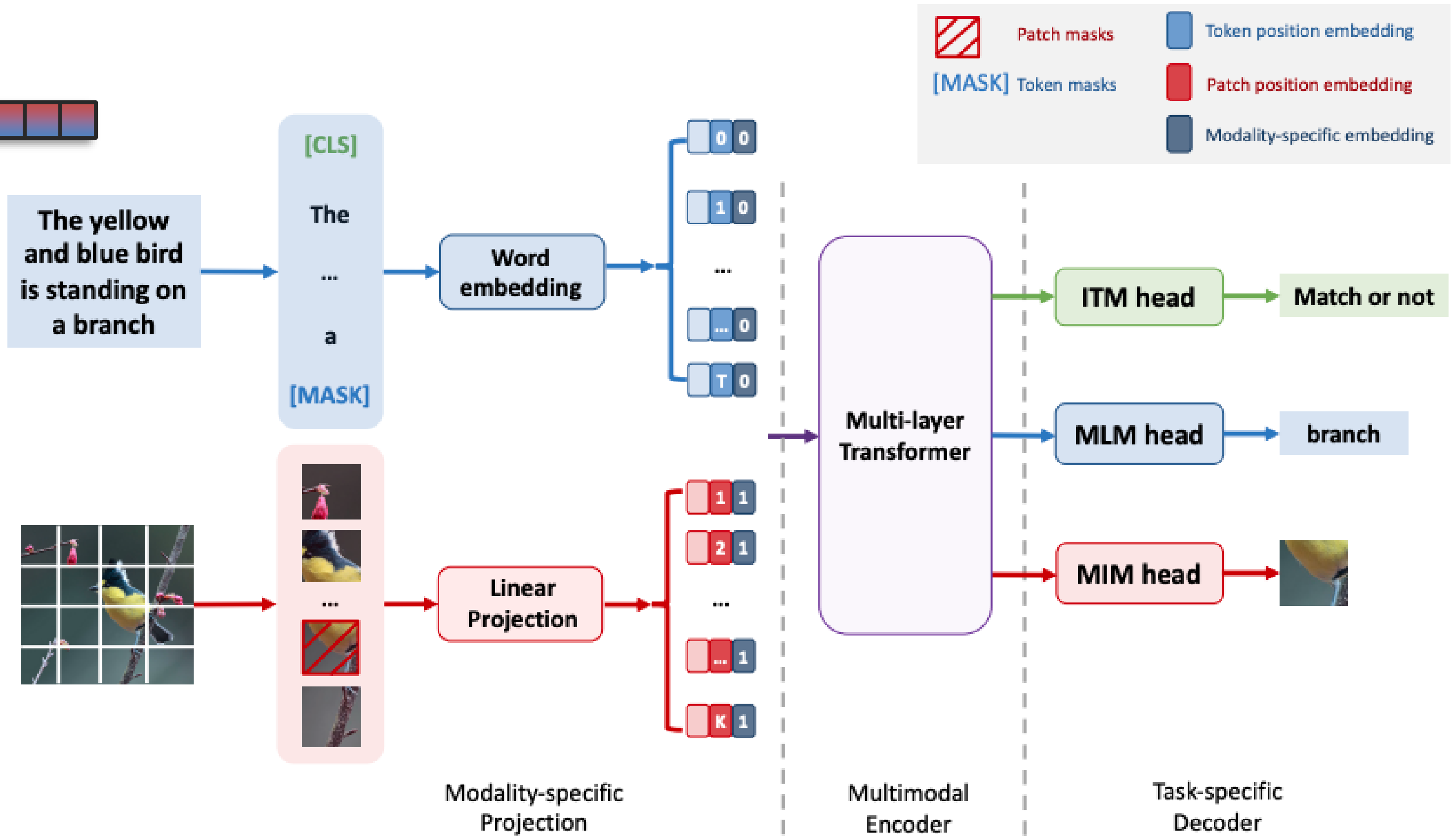
[Xue and Marculescu, Dynamic Multimodal Fusion, arxiv 2022]
[Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records. AAAI 2021]
[Liu et al., DARTS: Differentiable Architecture Search. ICLR 2019]

multisensory intelligence

Fusion with Heterogeneous Modalities

Example: From feature fusion to early fusion





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