Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 3.1: Multimodal Representation Fusion

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Upcoming Course Assignments

Week 3 Reading Assignment (due Friday 9/11 at 8pm ET)

- Two papers: a first about unimodal and a second about multimodal

Pre-proposal (due Wednesday 9/9 at 8pm ET)

- All team matching should be completed
- Describe your team’s dataset and present initial research ideas

Proposal, aka First Assignment (due Sunday 9/20 at 8pm ET)

- Literature review of your research problem
- Second revision of your research ideas (i.e., top-down approach)

Still looking for teammates?

See post on Piazza and add your follow-up post

Primary TAs

- Each team will have one primary TA
- Meetings with primary TA will be scheduled for next week
  - Feedback about 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

NEW: Asynchronous Lecture Questions on Piazza

- Questions live during the lecture are still the preferred approach
  - Ask them by raising you hand ☺

- If you prefer to ask asynchronously your question:
  - You can post them, as follow-discussion post, in Piazza
    - Up to 6 hours after the end of the lecture, if possible
  - With the help of the TA team, we plan to respond these questions asynchronously (within 24 hours, if possible)

Attendance using Poll Everywhere

1. Scan QR Code

   or

   https://pe.app/mmml

2. Enter your AndrewID

   (as your "name tag")

   Do not try to sign-in

3. Click "Share location"

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 3.1: Multimodal Representation Fusion

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

Lecture Objectives

- Vision Transformers
  - CNN with self-attention
  - Masked auto-encoder
- Multimodal representations
  - Cross-modal interactions
- Representation fusion
  - Additive and multiplicative fusion
  - Gated fusion
  - Fusion on raw modalities
  - Heterogeneity-aware fusion
- Measuring non-additive interactions

Vision Transformers

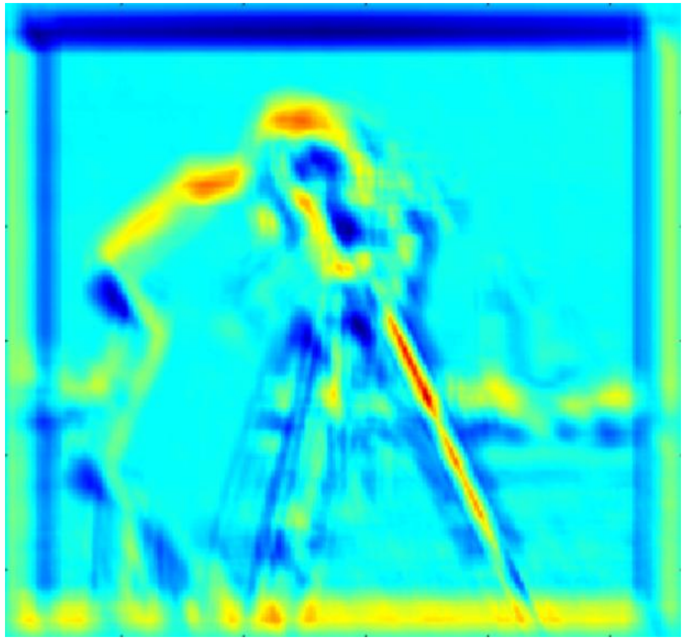
Convolution in 2D – Example

Input image

$*$

Convolution kernel

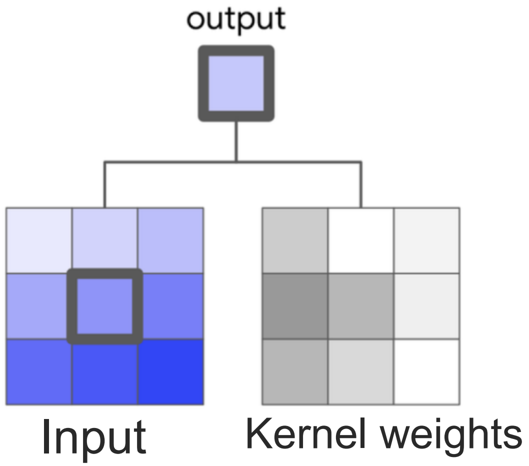
$=$



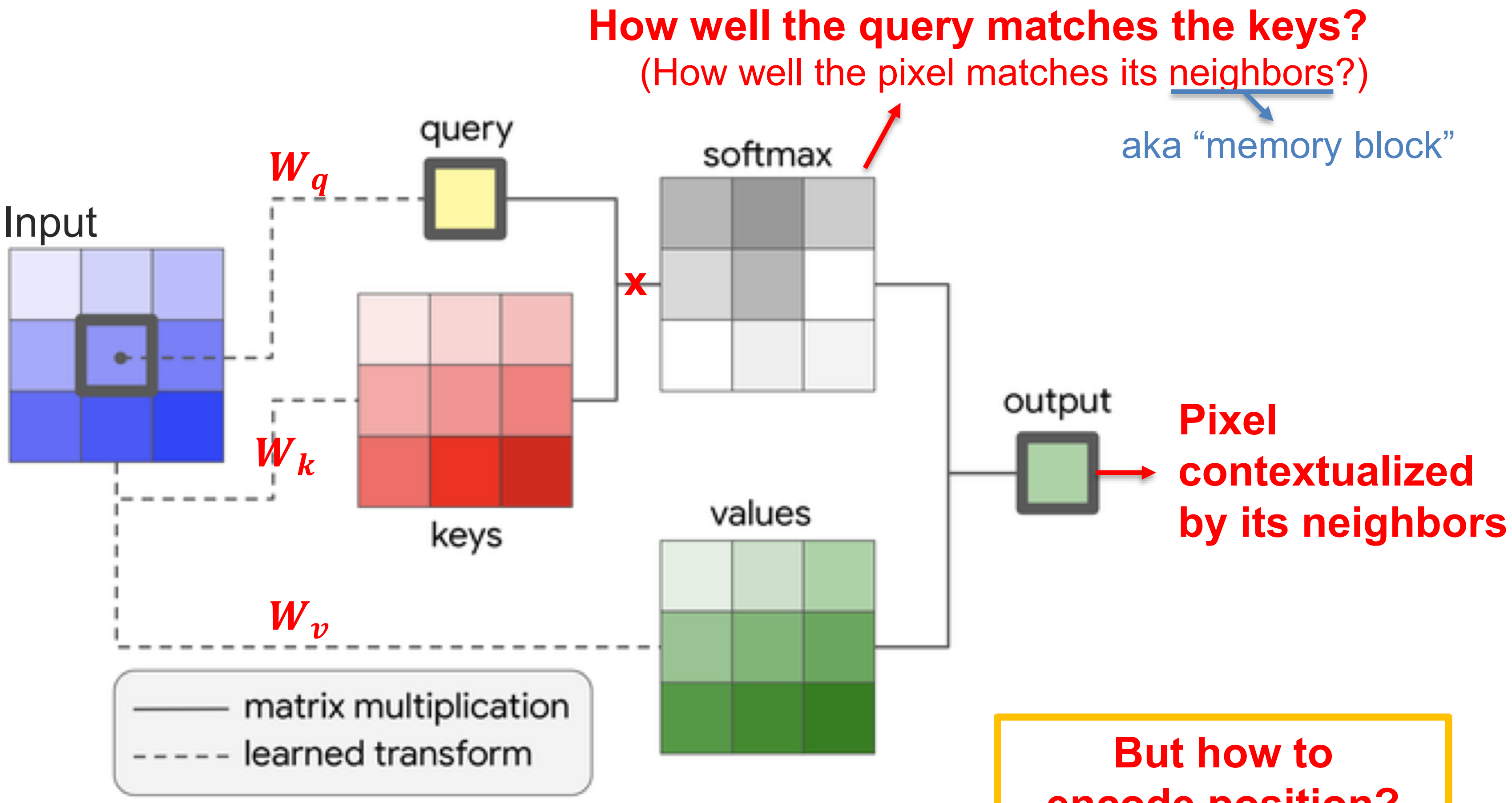
Response map

CNN

Replacing a CNN w/ Self-Attention



Convolution



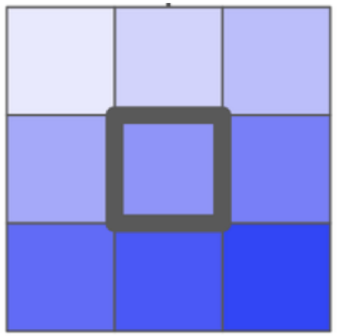
Self-attention

But how to encode position?

https://arxiv.org/abs/1906.05909

Replacing a CNN w/ Self-Attention

Image patch



2D relative position embedding

-1, -1	-1, 0	-1, 1	-1, 2
0, -1	0, 0	0, 1	0, 2
1, -1	1, 0	1, 1	1, 2
2, -1	2, 0	2, 1	2, 2

Position embedding is added to the key:

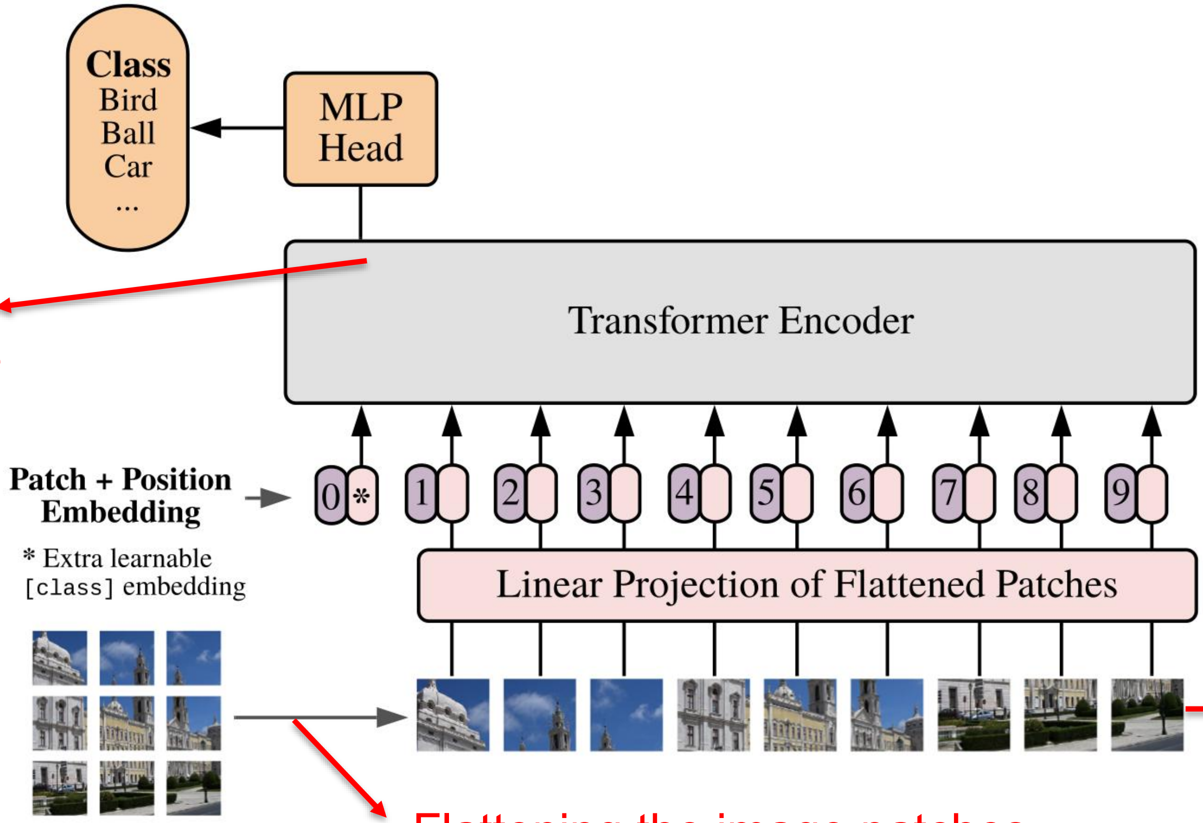
$$y_{ij} = \sum_{a,b \in \mathcal{N}_k(i,j)} \texttt{softmax}_{ab}\left(q_{ij}^{\top} k_{ab} + q_{ij}^{\top} r_{a-i,b-j}\right) v_{ab}$$

https://arxiv.org/abs/1906.05909

Vision Transformer (ViT)

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

Vision Transformer (ViT)



Embedding for the whole image

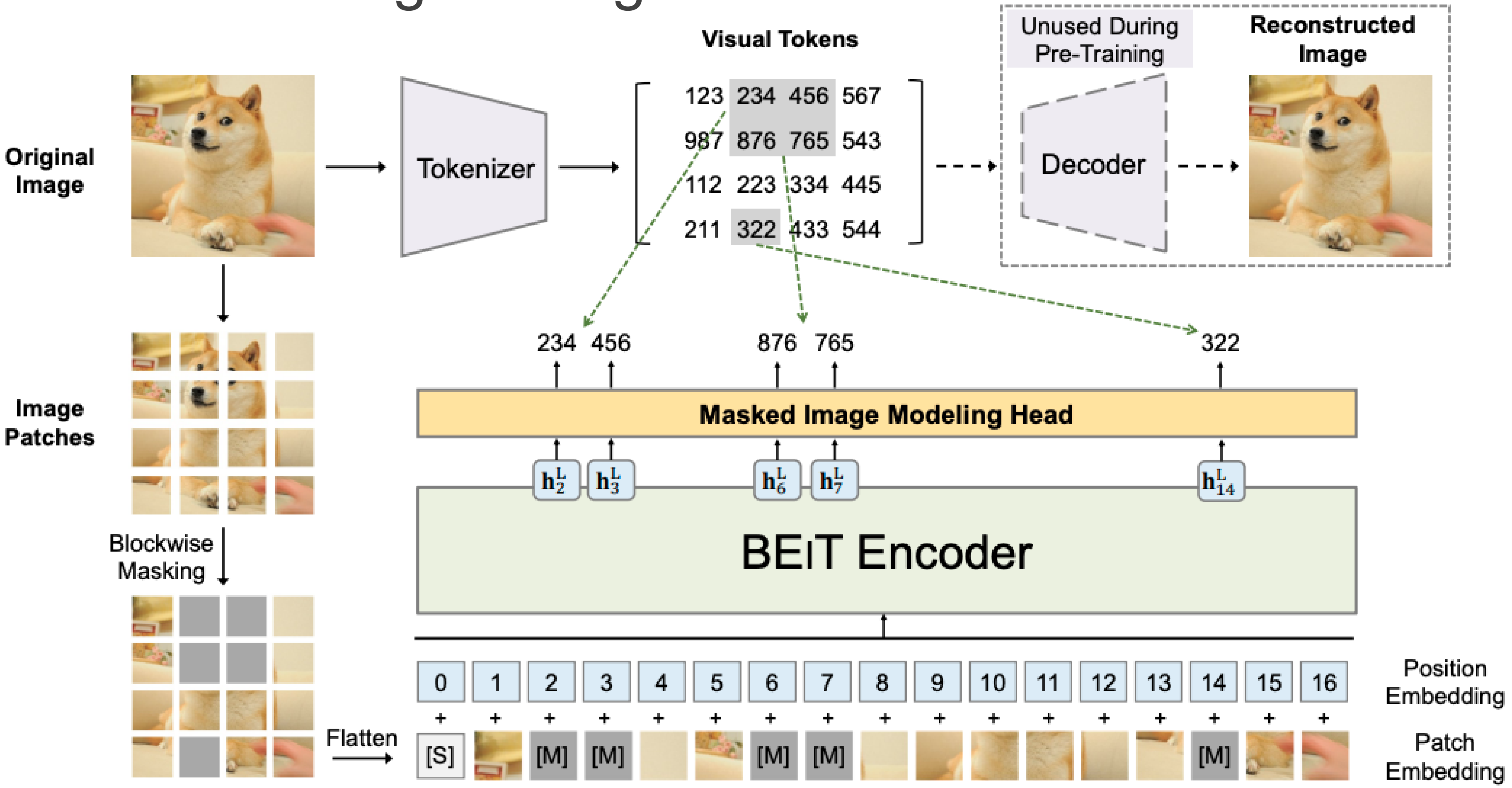
16x16 image patches

Flattening the image patches

Dosovitskiy, Alexey, et al. "An image is worth 16x16 words: Transformers for image recognition at scale." *arXiv* (2020).

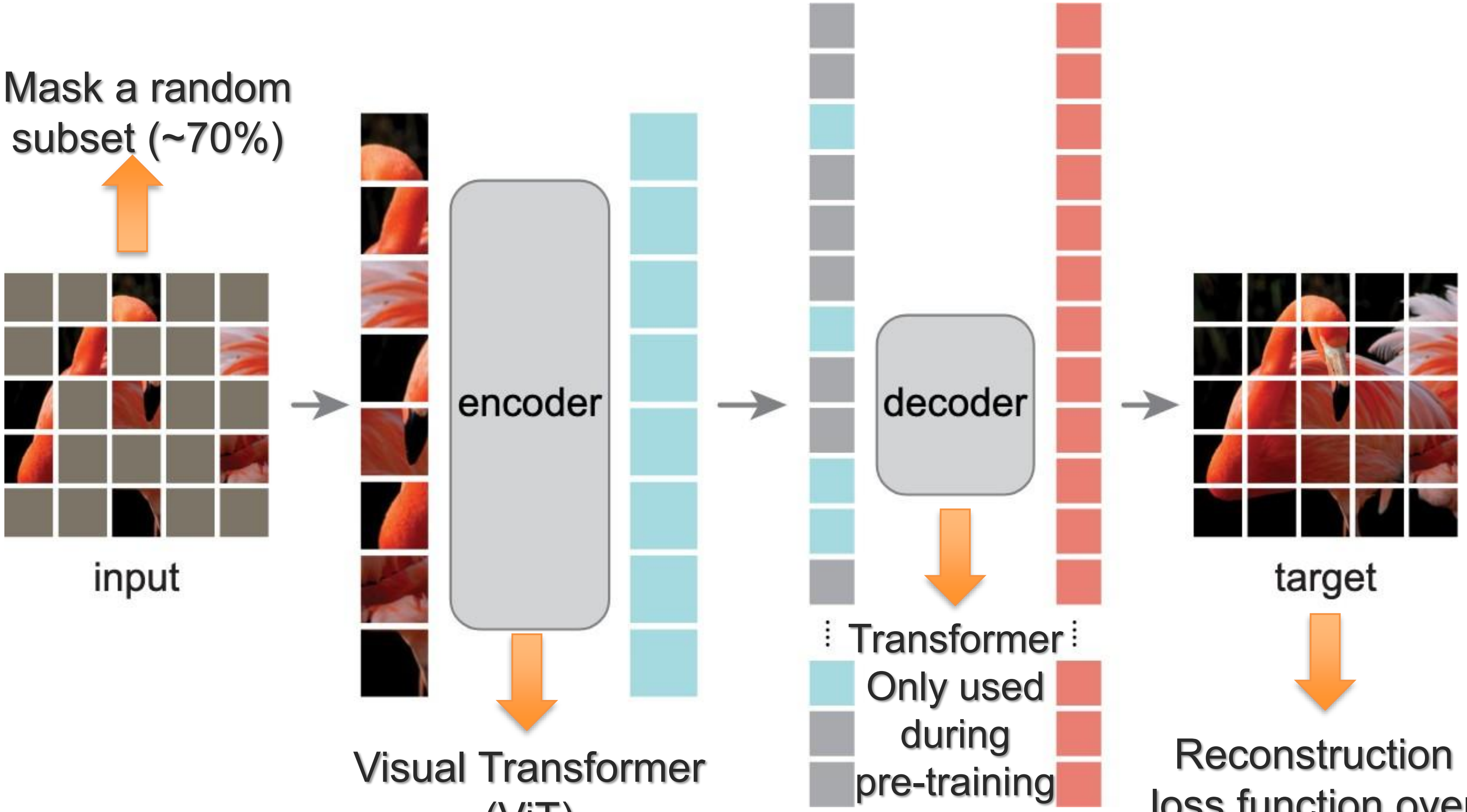
Visual Tokens

BeIT: BERT Pre-Training of Image Transformers



Masked Auto-Encoder (MAE)

Mask a random subset (~70%)



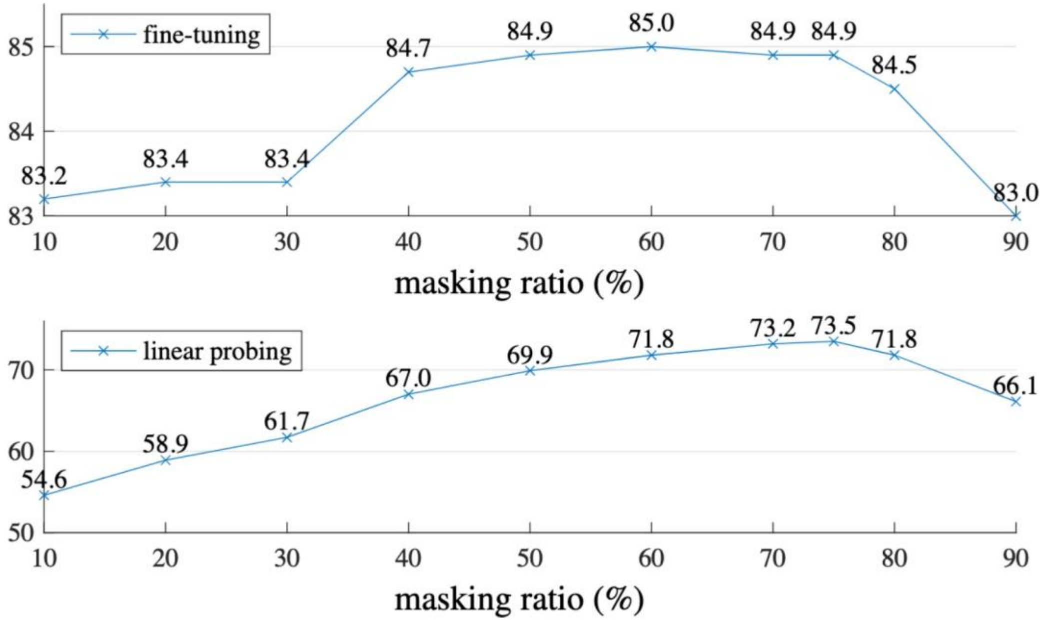
Visual Transformer (ViT)

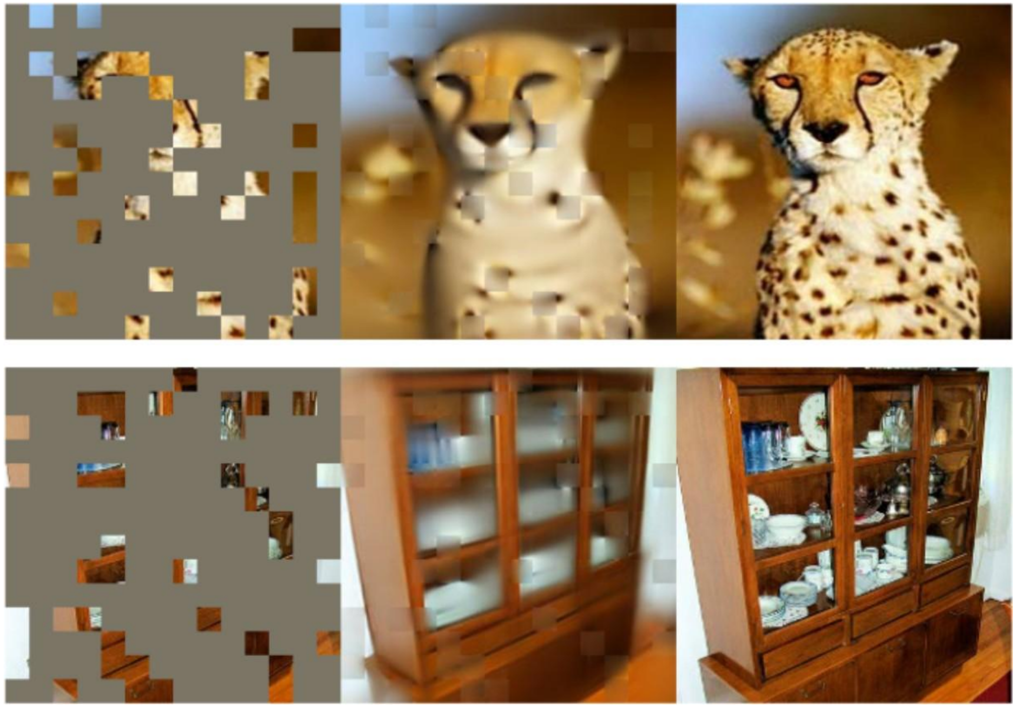
Transformer
Only used during pre-training

Reconstruction loss function over the whole image

He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

Masked Auto-Encoder (MAE)





He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022

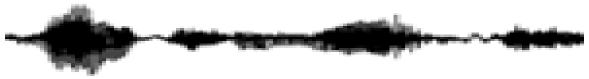
Multimodal Representation

Multimodal Machine Learning

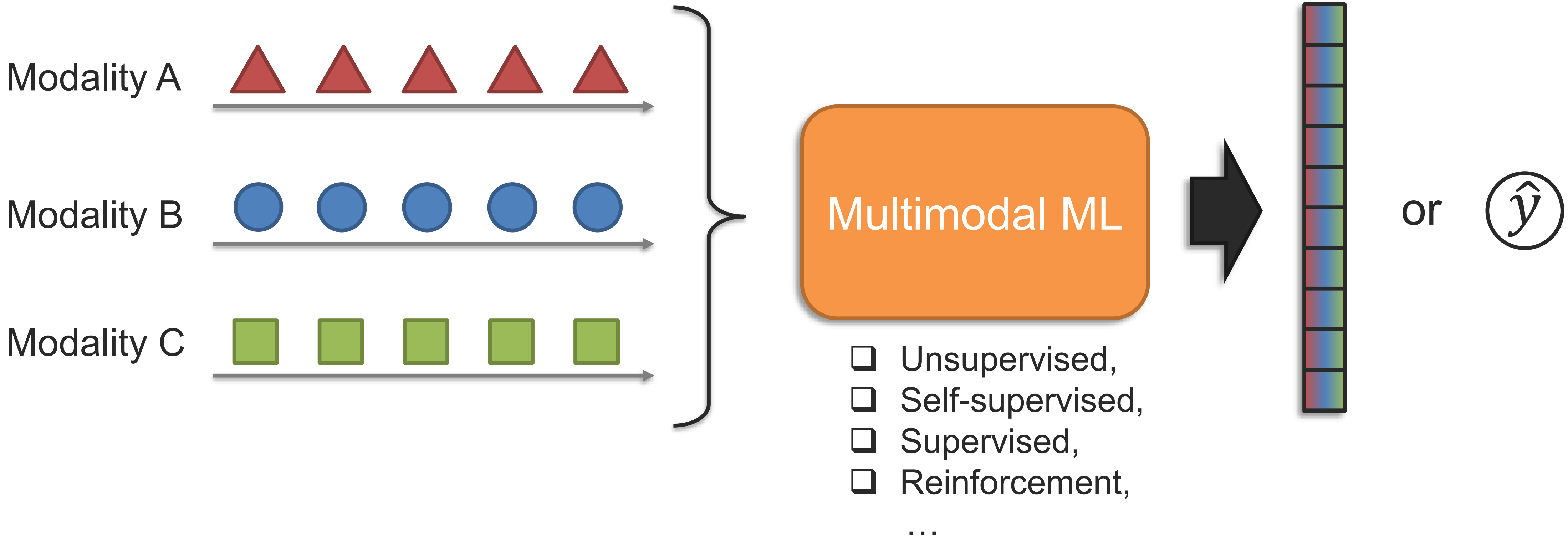
Language I really like this tutorial

Vision

Acoustic



Multimodal Machine Learning

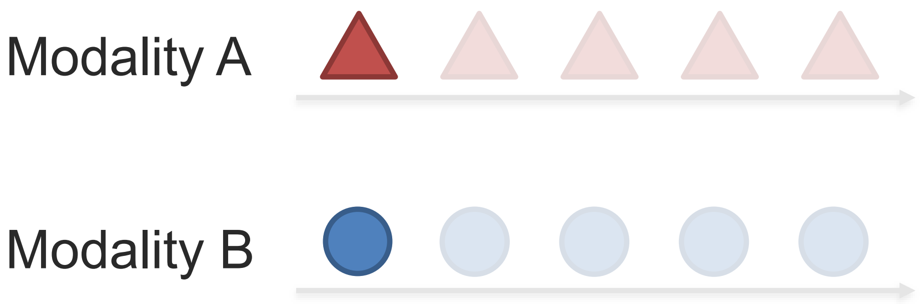


Challenge 1: Representation

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

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

Individual elements:



It can be seen as a "local" representation

or

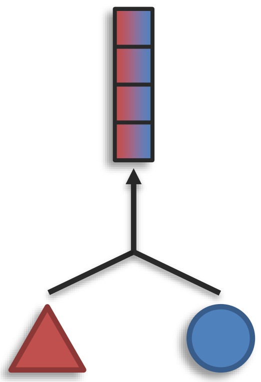
representation using holistic features

Challenge 1: Representation

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

Sub-challenges:

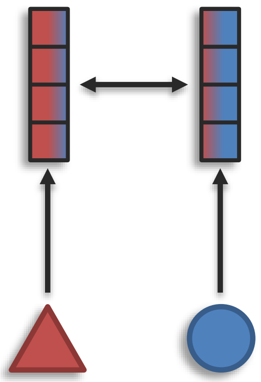
Fusion



modalities **>** # representations

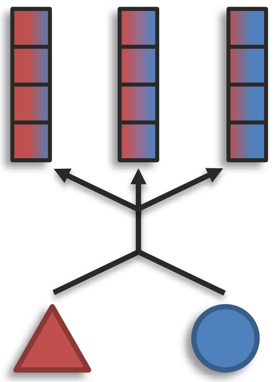
Today

Coordination



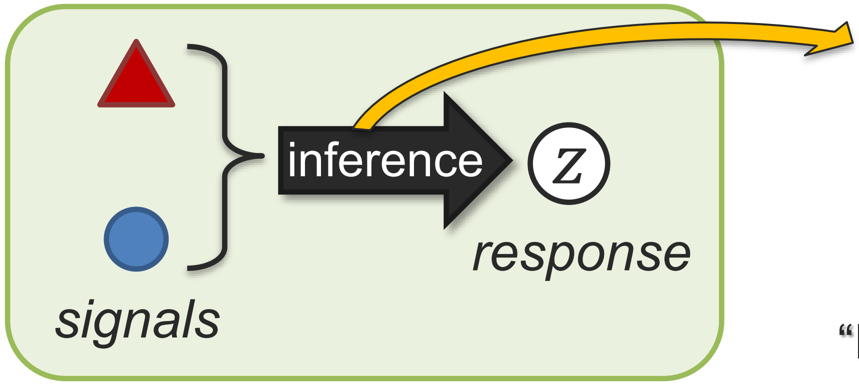
modalities **=** # representations

Fission



modalities **<** # representations

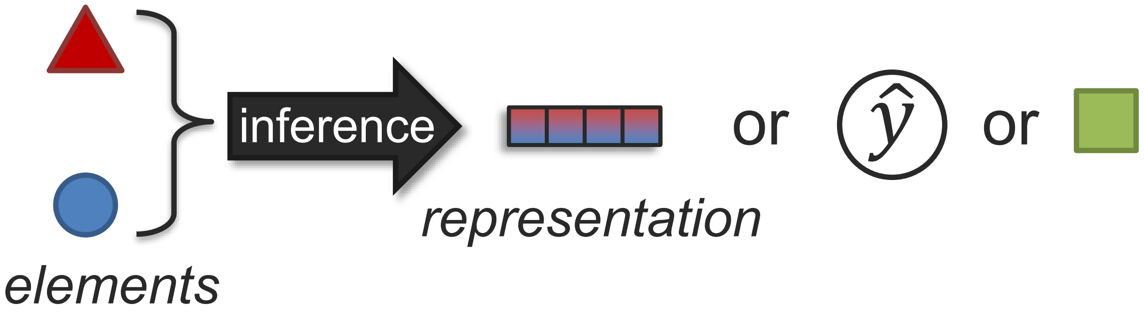
Cross-modal Interactions



Interactions happen during inference!

"Inference" examples:

- Representation fusion
- Prediction task
- Modality translation



Interacting Modalities

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

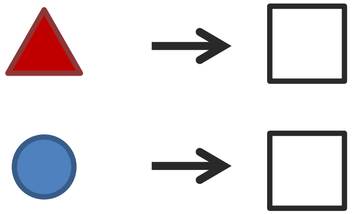
Is this indoors?

inference → *Yes!*

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

inference → *Yes!*

Redundant



Interacting Modalities

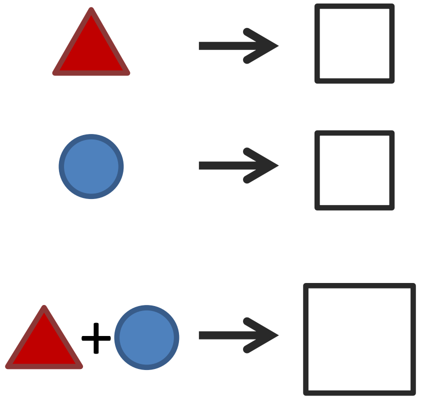
Interactions: How multimodal information changes when modalities are combined for a response.

Is this indoors?

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

inference → ***Yes!***

Redundant



Enhancement

Interacting Modalities

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

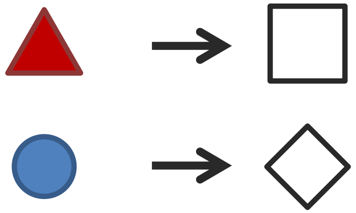
Is this a living room?

inference → *Yes!*

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

inference → *No, probably study room.*

Non-redundant



Interacting Modalities

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

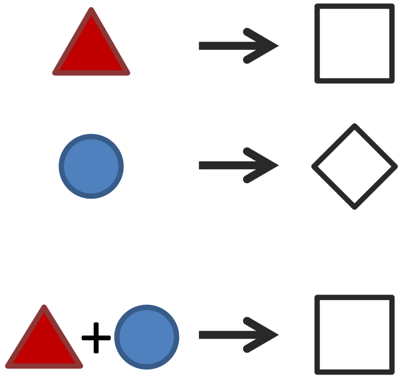
Is this a living room?

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

inference

Yes!

Non-redundant



Dominance

Interacting Modalities

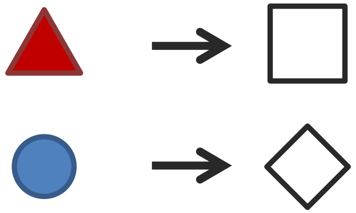
Interactions: How multimodal information changes when modalities are combined for a response.

Should I work here?

inference

Maybe? Comfy sofa but table's too small.

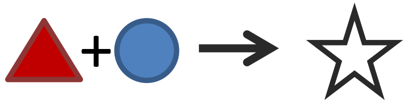
Non-redundant



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

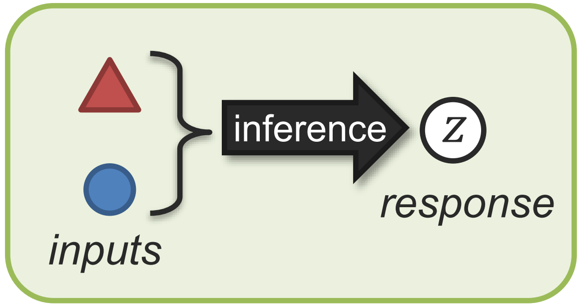
inference

Maybe? Clean and there's tea.



Emergence

Taxonomy of Interaction Responses – A Behavioral Science View

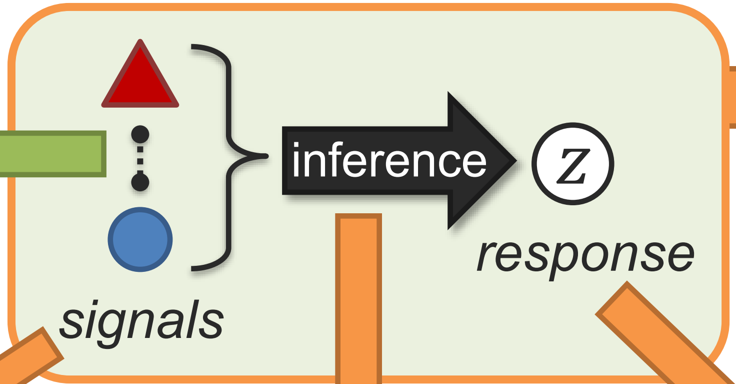


	signal	response	signal	response	
Redundancy	a →	□	a+b →	□	Equivalence
	b →	□	a+b →	☐ (larger)	Enhancement
Nonredundancy	a →	□	a+b →	□ and ○	Independence
	b →	○	a+b →	□	Dominance
			a+b →	☐ (larger) (or □)	Modulation
			a+b →	△	Emergence

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

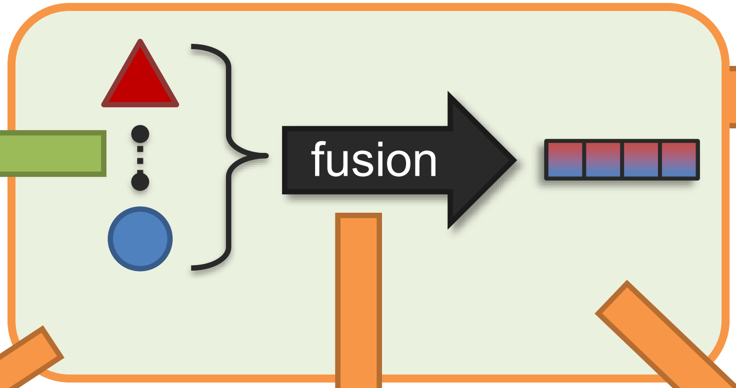
…

1 Responses

- Redundancy
- Non-redundancy
- Dominance
- Emergence

…

Cross-modal Interactions – Representation Fusion



Next lecture

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
- …

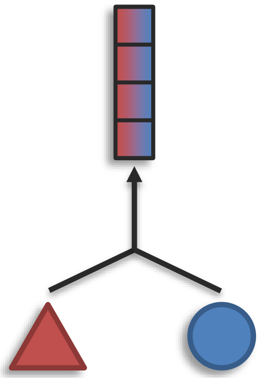
Today

1 Responses

- Redundancy
- Non-redundancy
- Dominance
- Emergence
- …

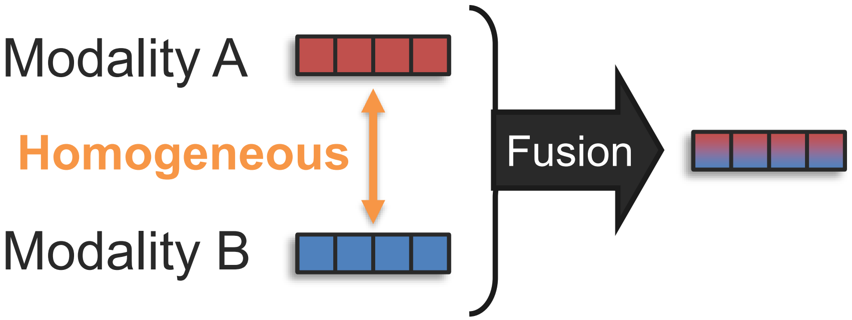
Representation Fusion

Sub-Challenge 1a: Representation Fusion

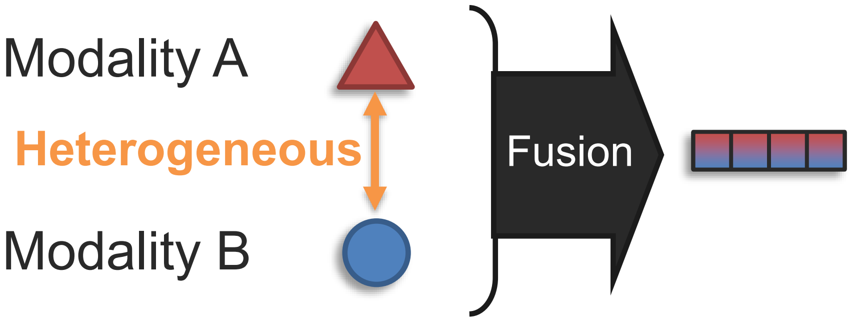


Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

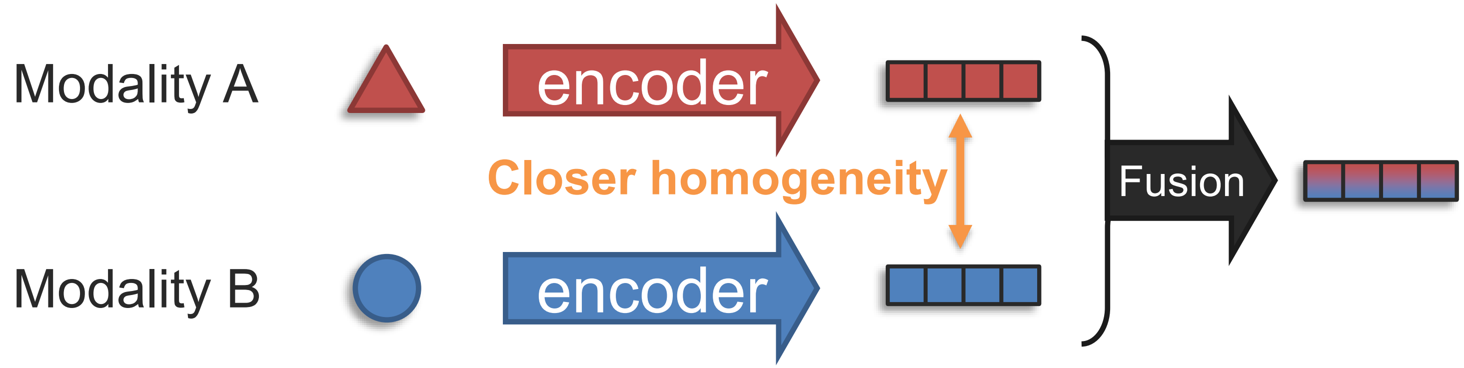
Basic fusion:



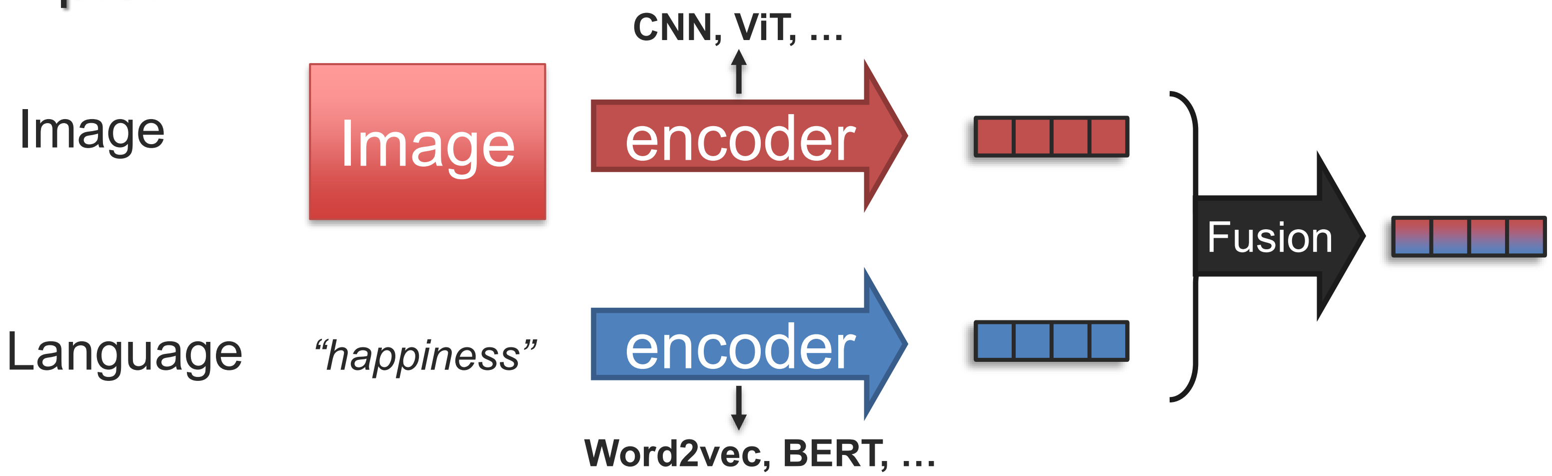
Raw-modality fusion:



Fusion with Unimodal Encoders



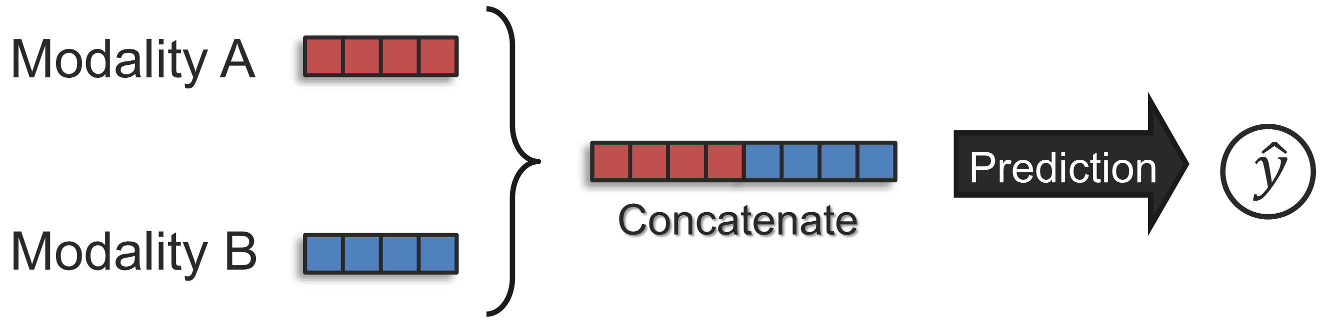
Example:



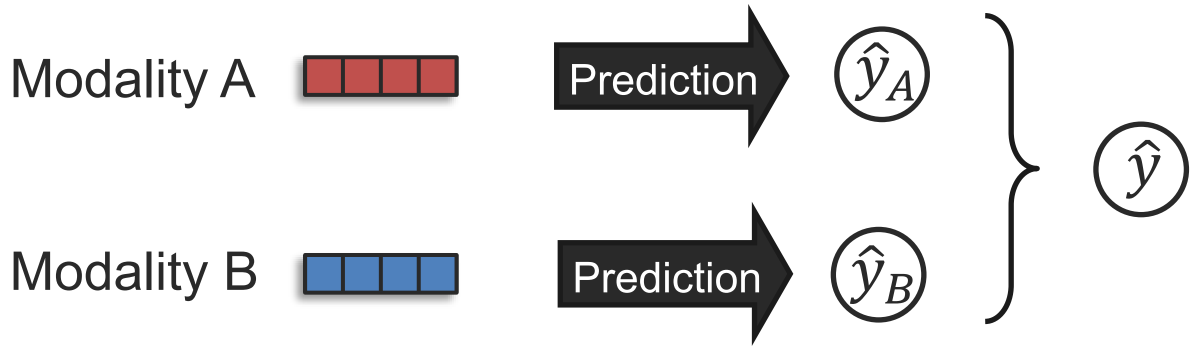
➡ Unimodal encoders can be jointly learned with fusion network, or pre-trained

Early and Late Fusion – A historical View

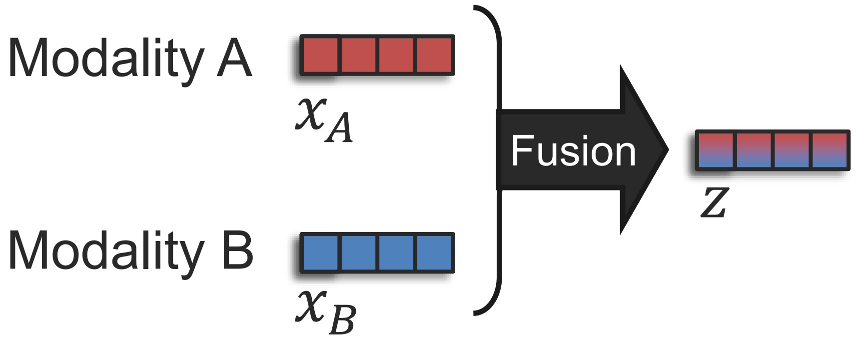
Early fusion:



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)

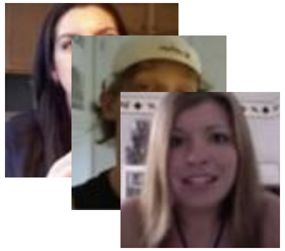
Linear regression:

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

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:

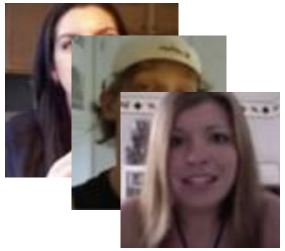
$$y = \underbrace{w_0}_{\text{intercept (bias term)}} + \underbrace{w_1 x_A + w_2 x_B}_{\text{Additive terms}} + \underbrace{w_3 (x_A \times x_b)}_{\text{Multiplicative term}} + \underbrace{\epsilon}_{\text{error (residual term)}}$$

- 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)

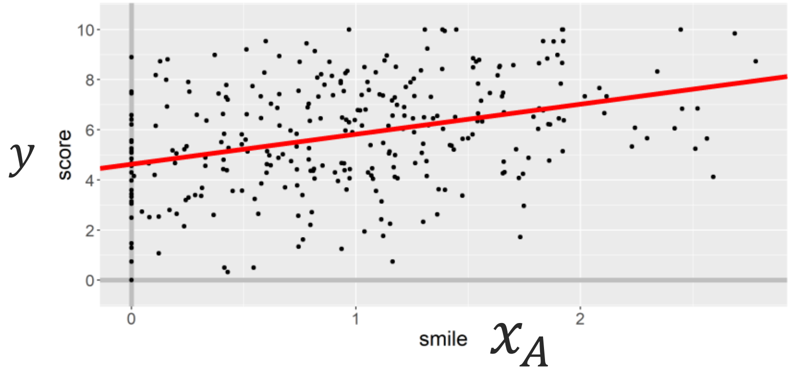
H1: *Does smiling reveal what the audience score was?*

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	$[0.83, 1.57]$

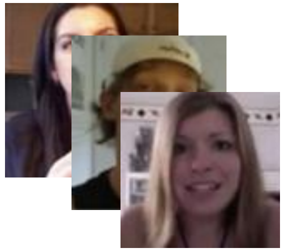
Confidence interval does not contain 0, so effect is significant

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

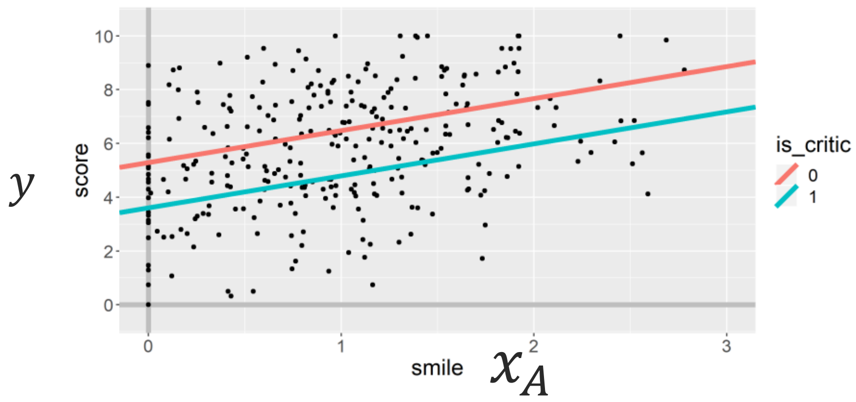
x_B: professional status (0=non-critic, 1=critic)

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]

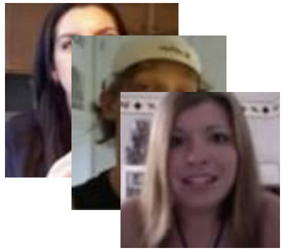
Positive effect

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)

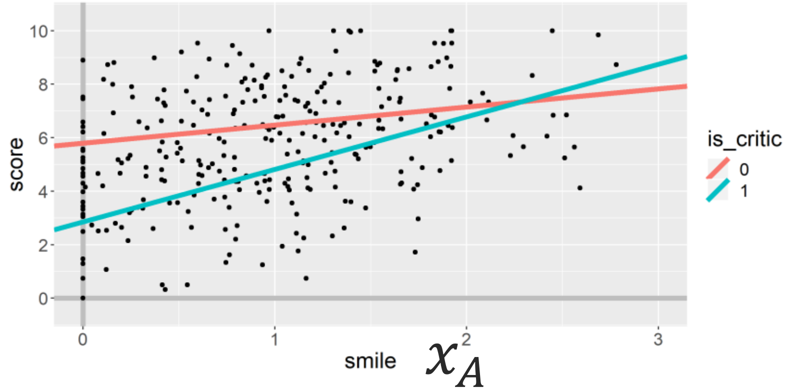
H1: *Does smiling reveal what the audience score was?*

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

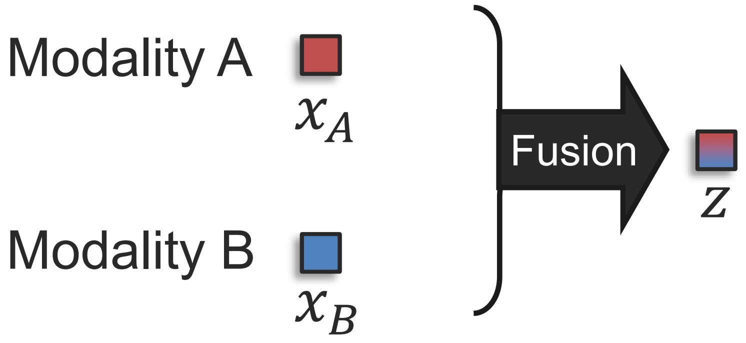


x_A

	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:

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

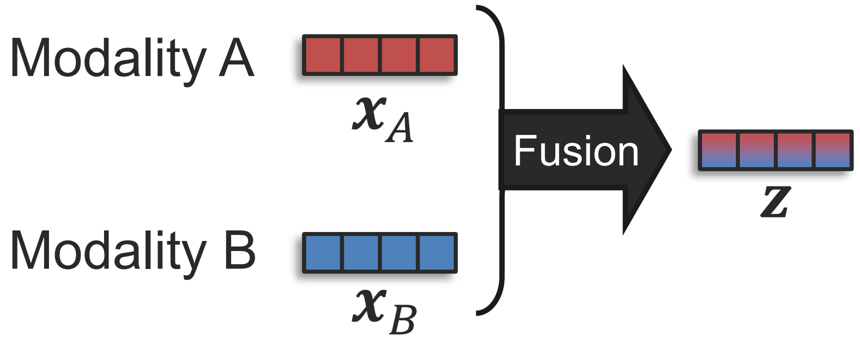
3. Additive and multiplicative terms:

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

Additive Fusion

Back to multivariate case!

(multi-dimensional features)

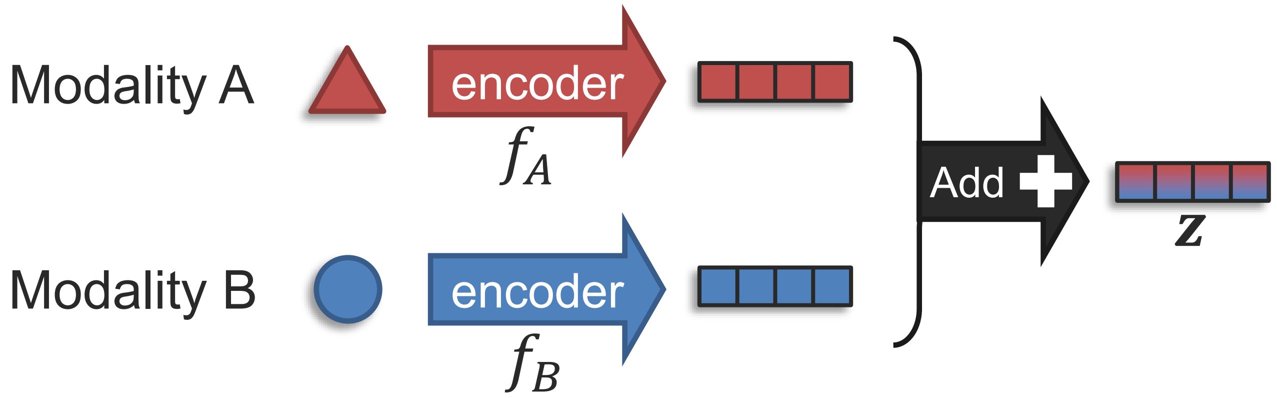


Additive fusion:

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

1-layer neural network can be seen as additive

With unimodal encoders:



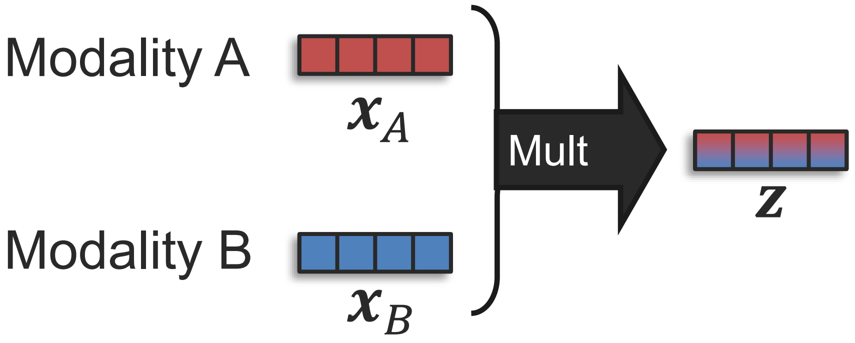
Additive fusion:

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

It could be seen as an ensemble approach

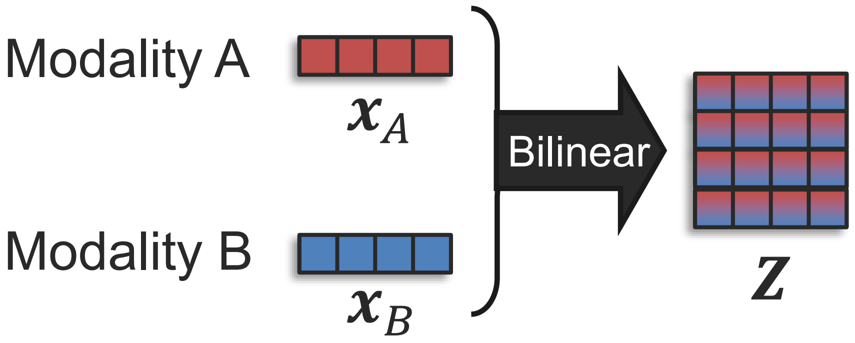
(late fusion)

Multiplicative Fusion



Simple multiplicative fusion:

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

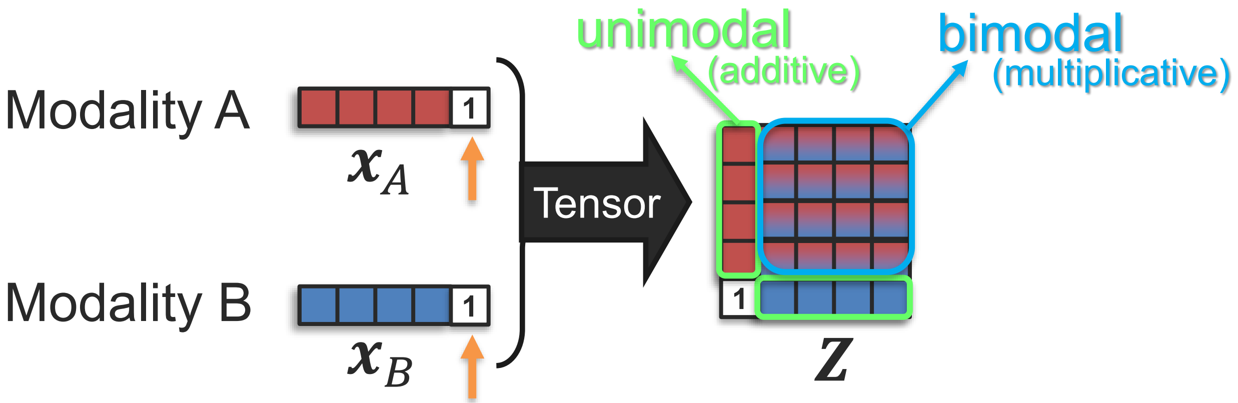


Bilinear Fusion:

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

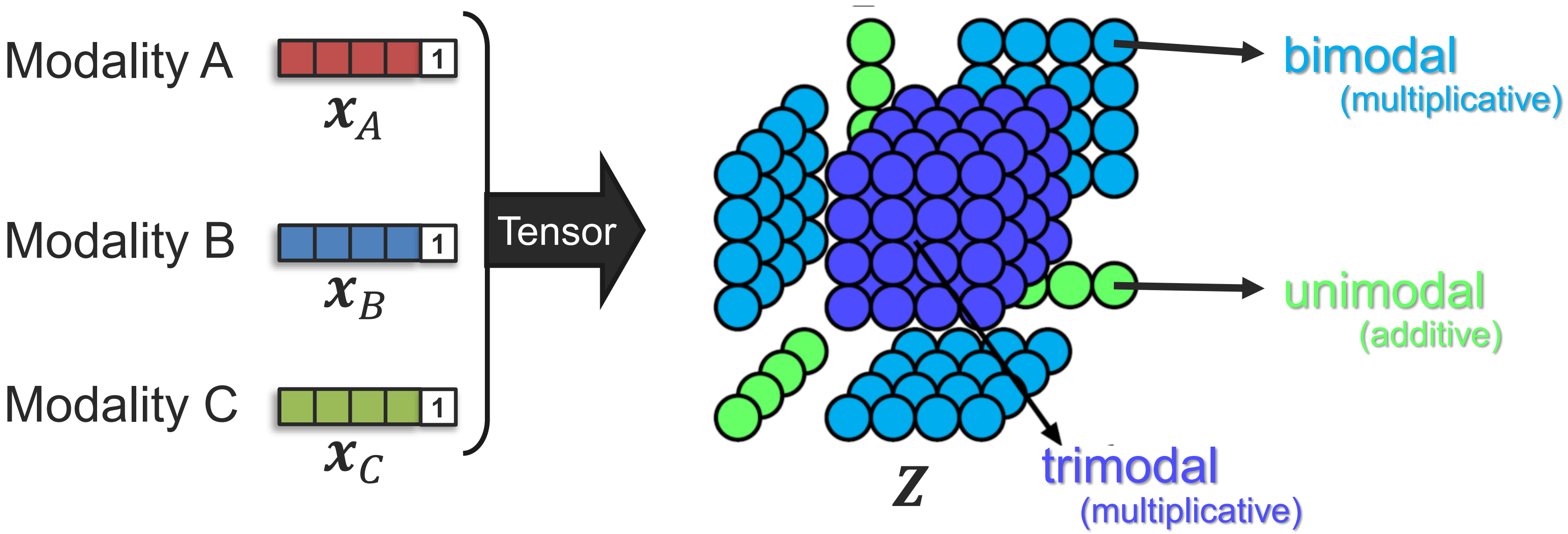
Jayakumar et al., Multiplicative Interactions and Where to Find Them, ICLR 2020

Tensor Fusion



Tensor Fusion (bimodal):

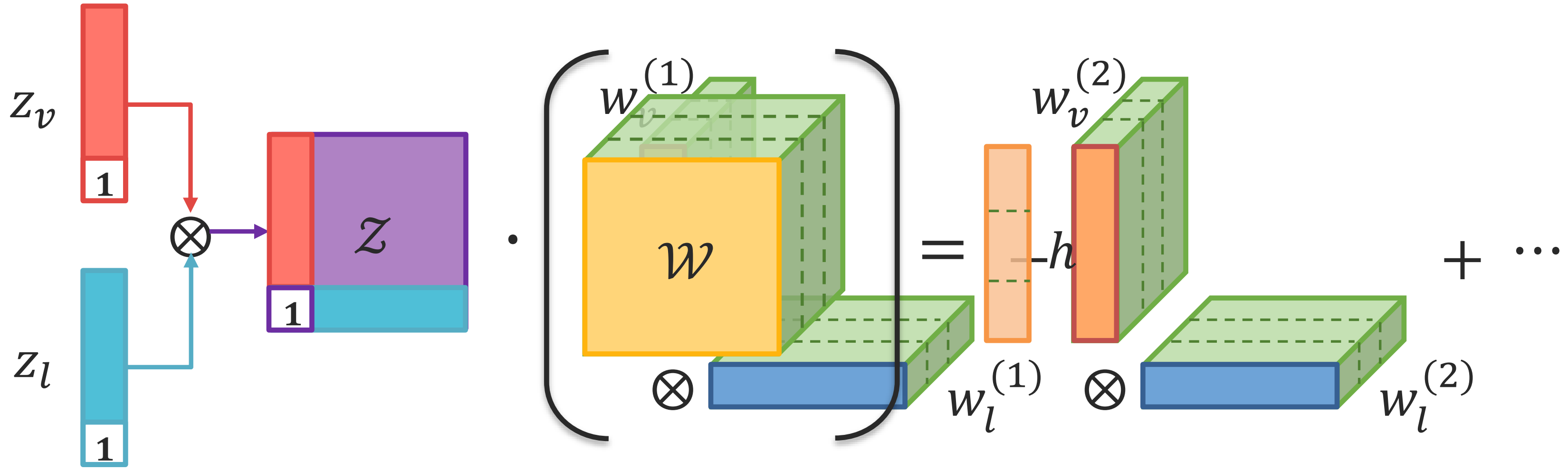
$$Z = w([x_A \quad 1]^T \cdot [x_B \quad 1])$$



… but the weight matrix may end up quite large!

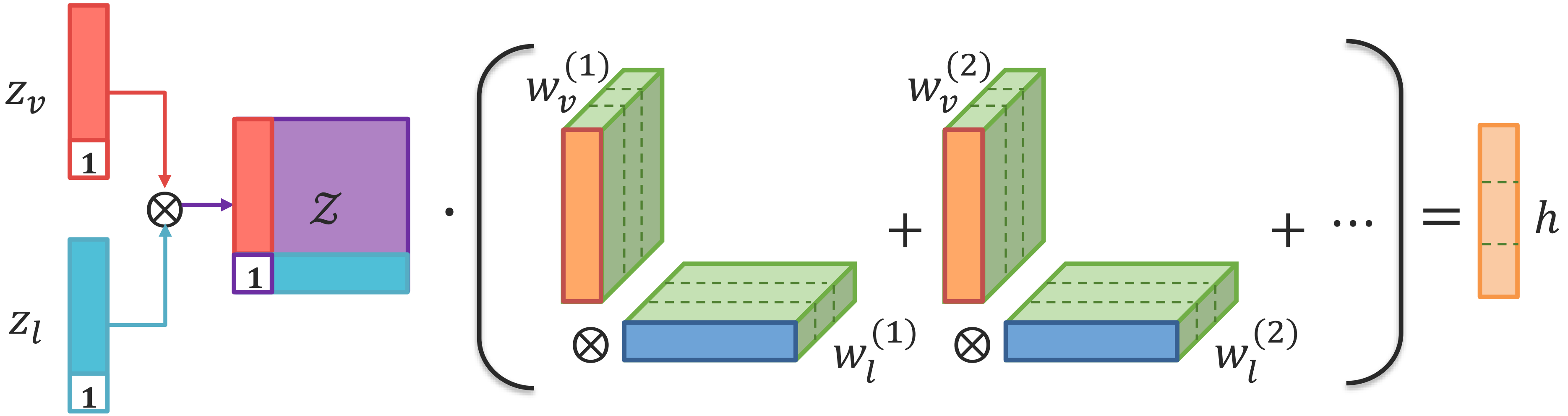
Zadeh et al., Tensor Fusion Network for Multimodal Sentiment Analysis, EMNLP 2017

Low-rank Fusion



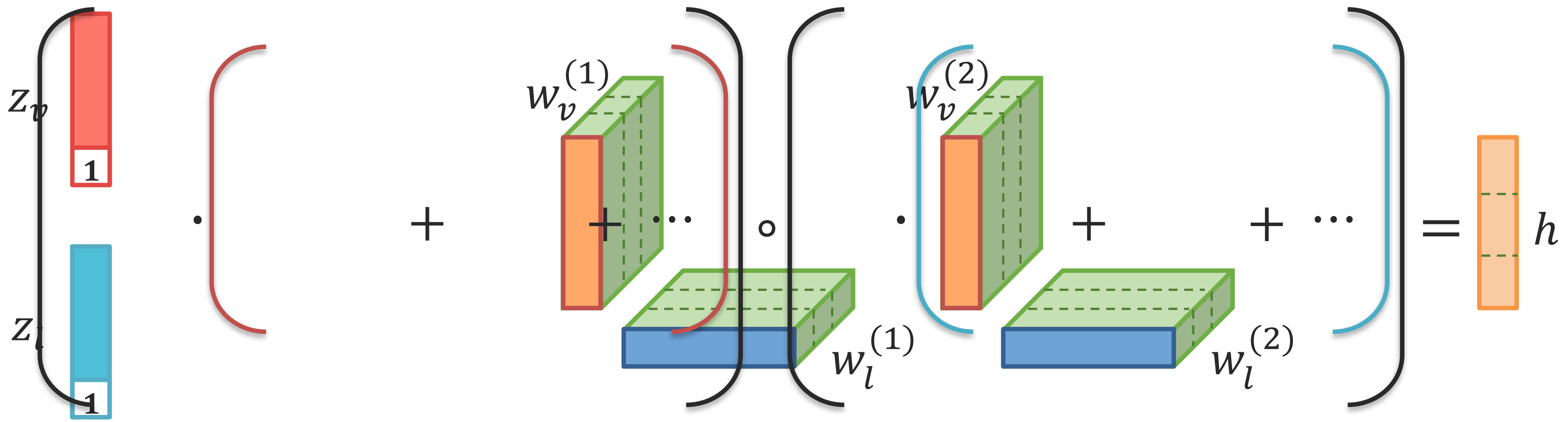
Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

Low-rank Fusion



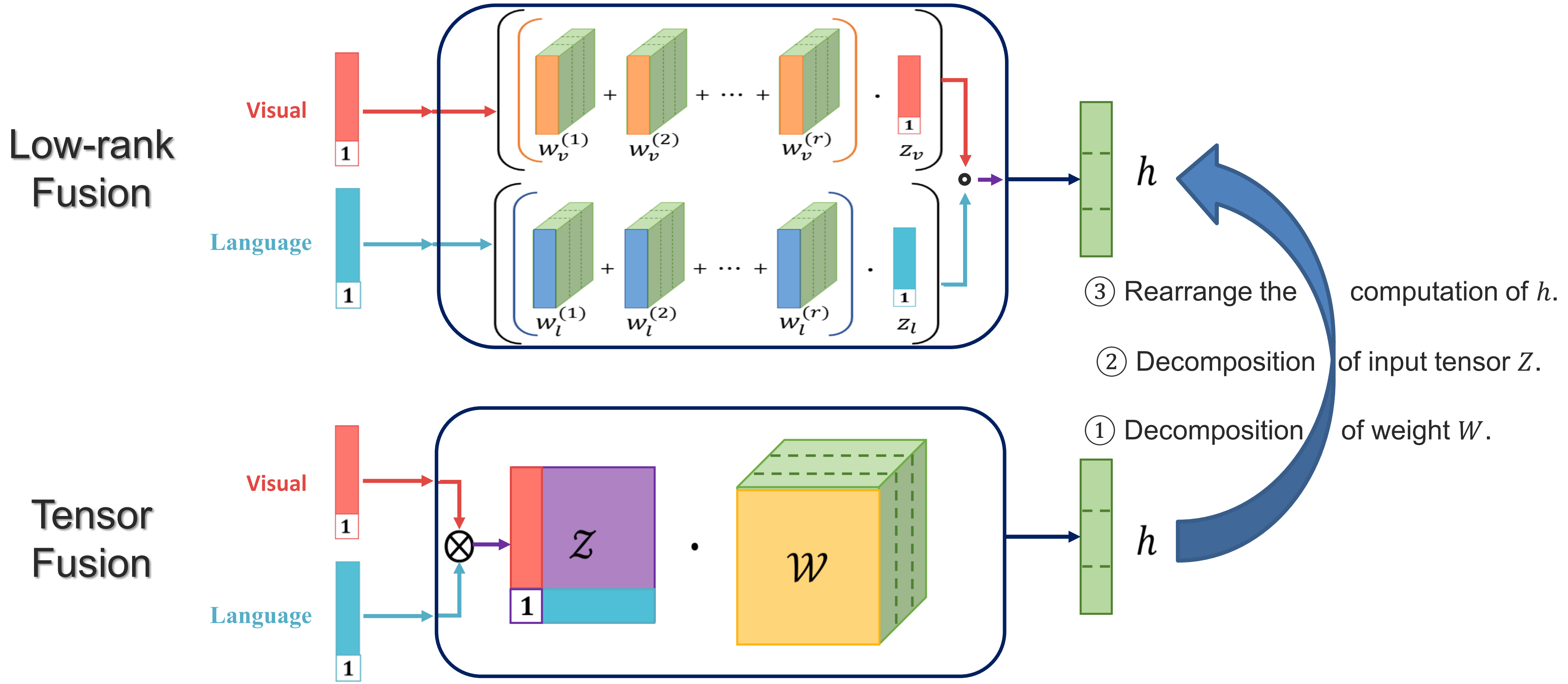
Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

Low-rank Fusion



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

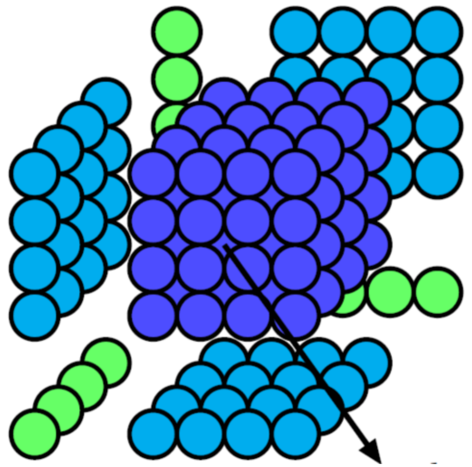
Low-rank Fusion



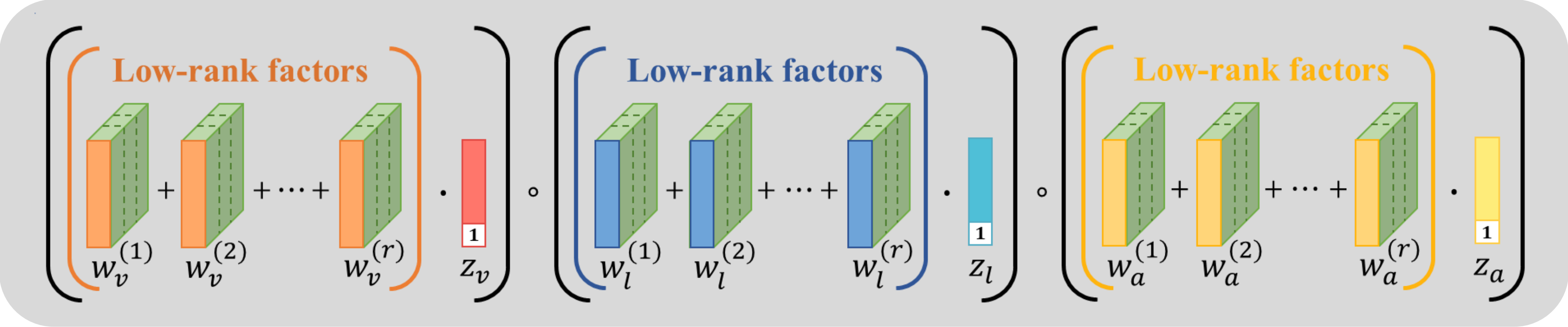
Liu et al., Efficient Low-rank Multimodal Fusion with Modality-Specific Factors, ACL 2018

Low-rank Fusion with Trimodal Input

Tensor Fusion



Low-rank Fusion :



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):

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

Can we add higher-order interaction terms?

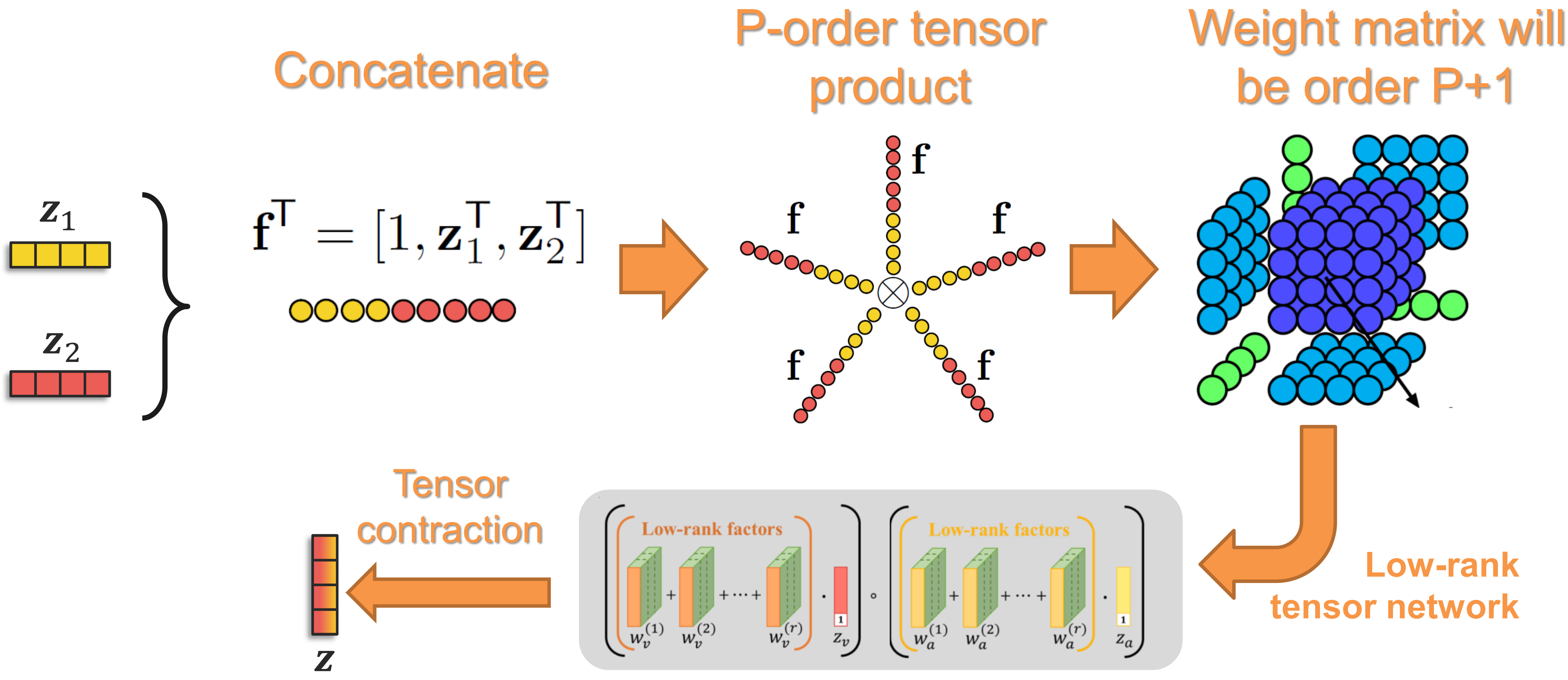
For example:

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

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

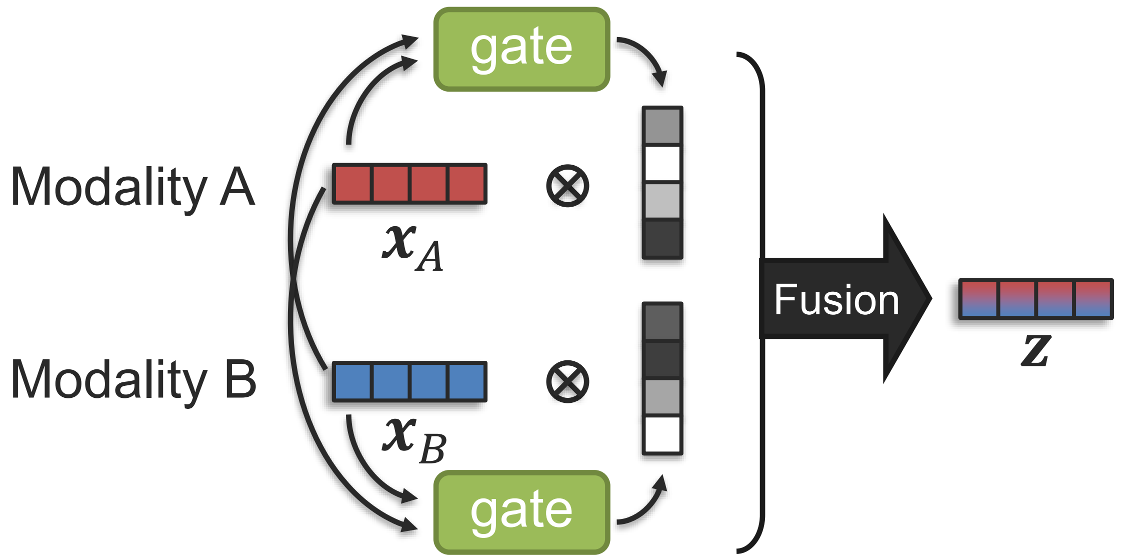
$$+w_{10} (x_B^3 \times x_C^3)$$

High-Order Polynomial Fusion



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

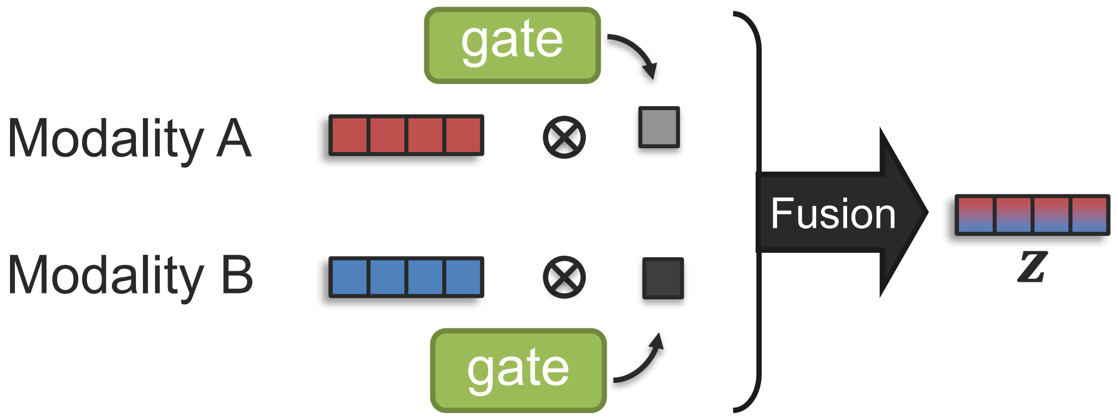
Gated Fusion



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 →

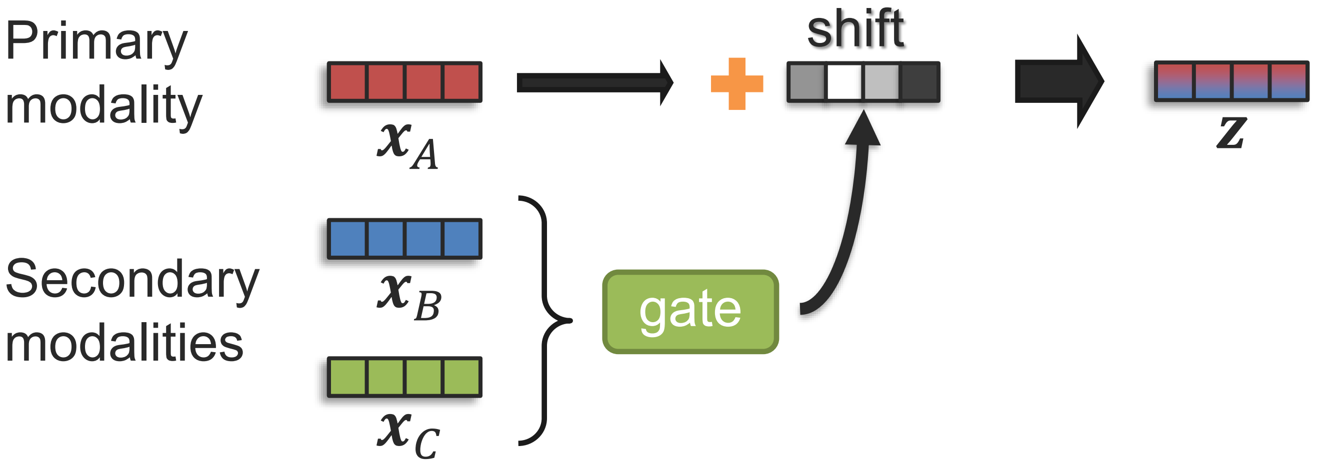
Easier to compute derivative (gradient)

Hard attention gate →

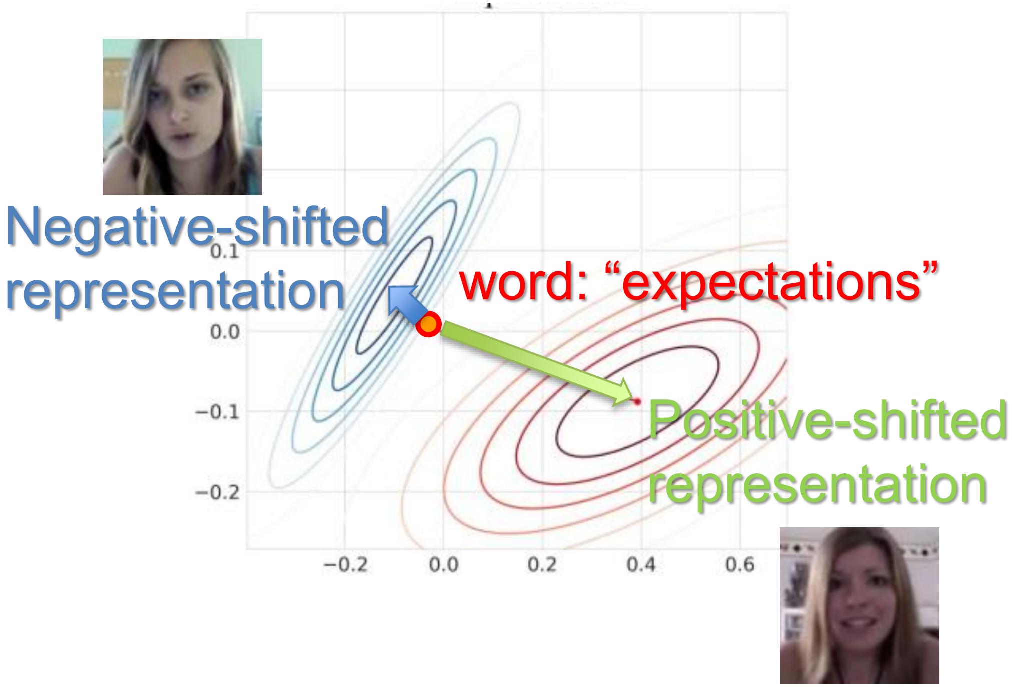
Derivative is harder (e.g., use reinforcement learning)

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

Modality-Shifting Fusion



Example with language modality:

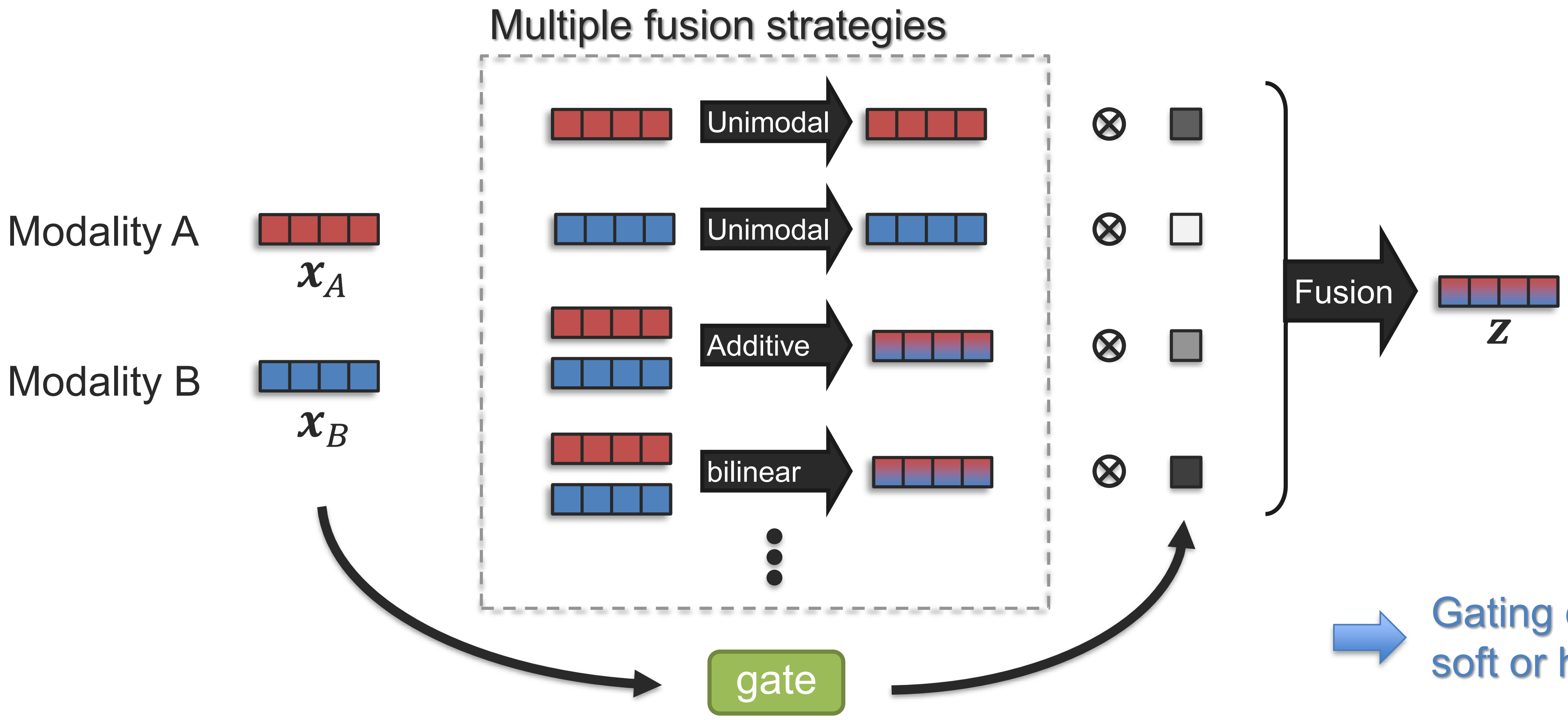


Primary modality: language

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]

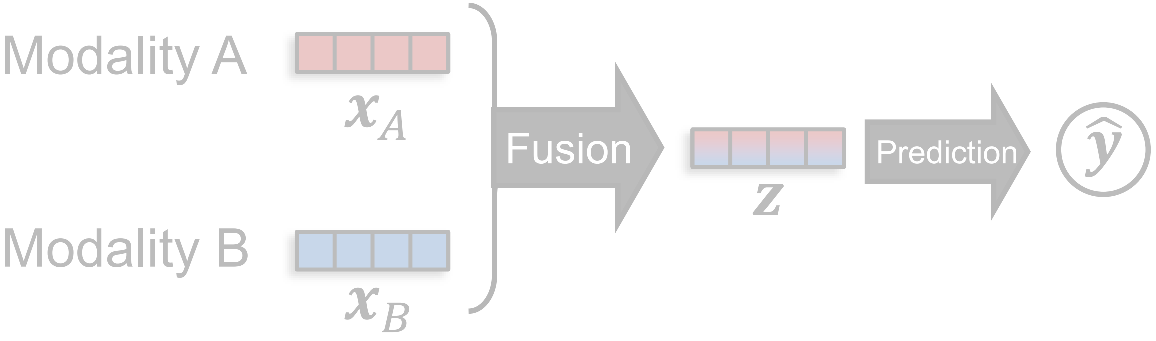
Mixture of Fusions

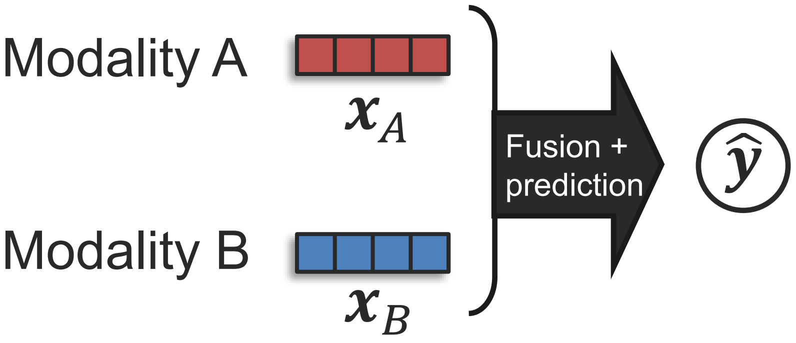


Gating can be with soft or hard attention

[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:

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

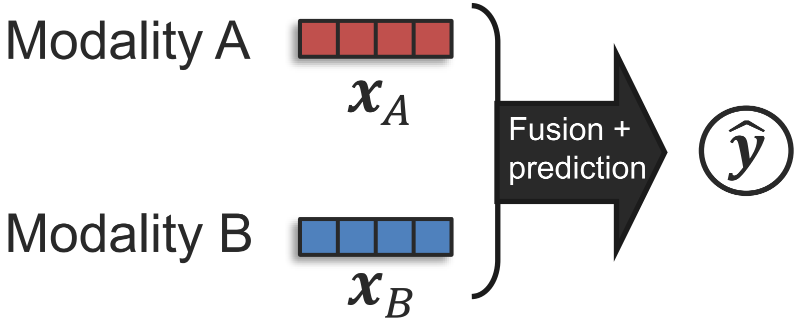
where f could be a multi-layer perceptron or any nonlinear model

➡ 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



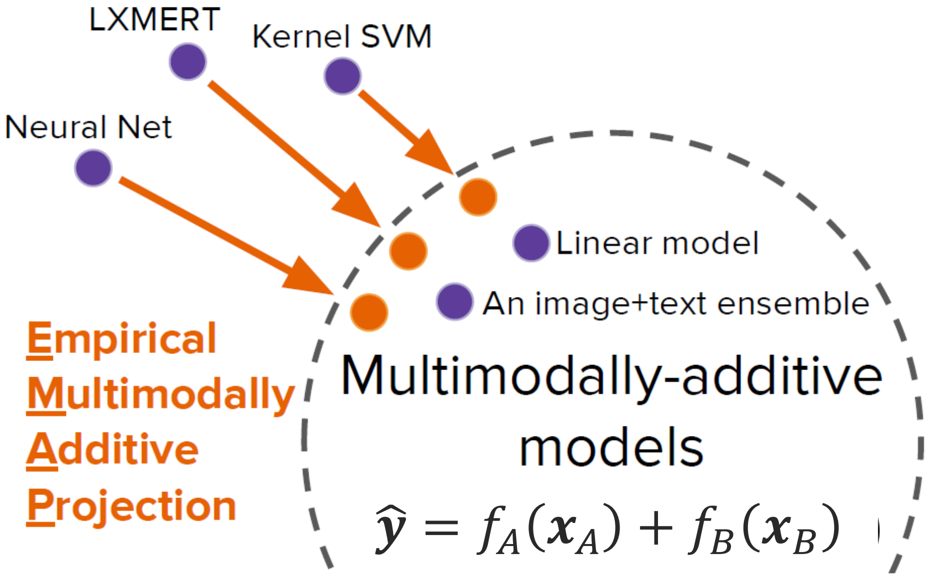
Nonlinear fusion:

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

Projection?

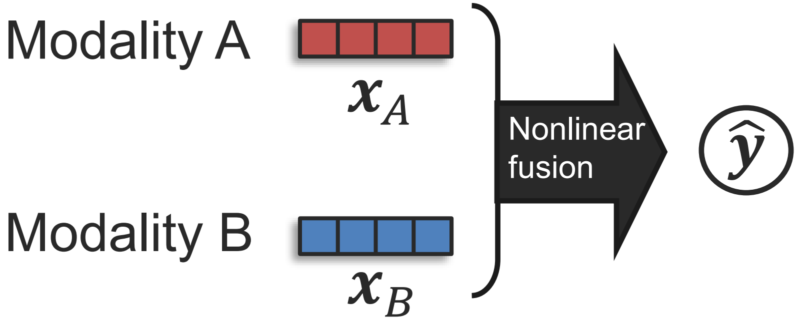
Additive fusion:

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



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):

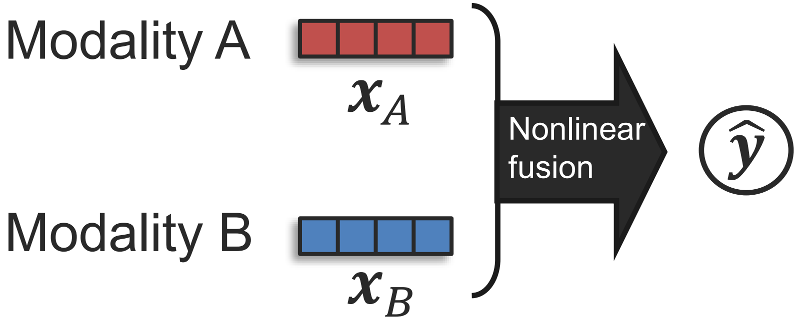
$$\tilde{f}(\boldsymbol{x}_A, \boldsymbol{x}_B) = \underbrace{\mathbb{E}_{\boldsymbol{x}_B}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_A(\boldsymbol{x}_A)} + \underbrace{\mathbb{E}_{\boldsymbol{x}_A}[f(\boldsymbol{x}_A, \boldsymbol{x}_B)]}_{f_B(\boldsymbol{x}_B)}$$

Modality A + Modality B

Additive fusion
(approximation)

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

Measuring Non-Additive Interactions



Nonlinear fusion:

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

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]

Learning Non-additive Bimodal and Trimodal Interactions

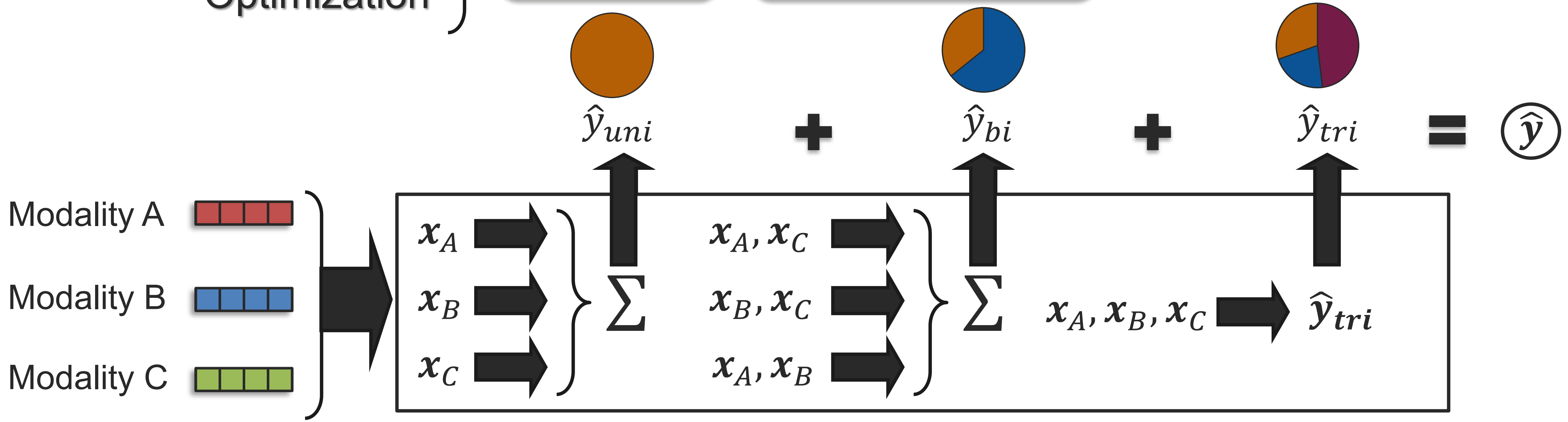
Idea: prioritize simpler interactions

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

residual | residual

Multimodal Residual Optimization

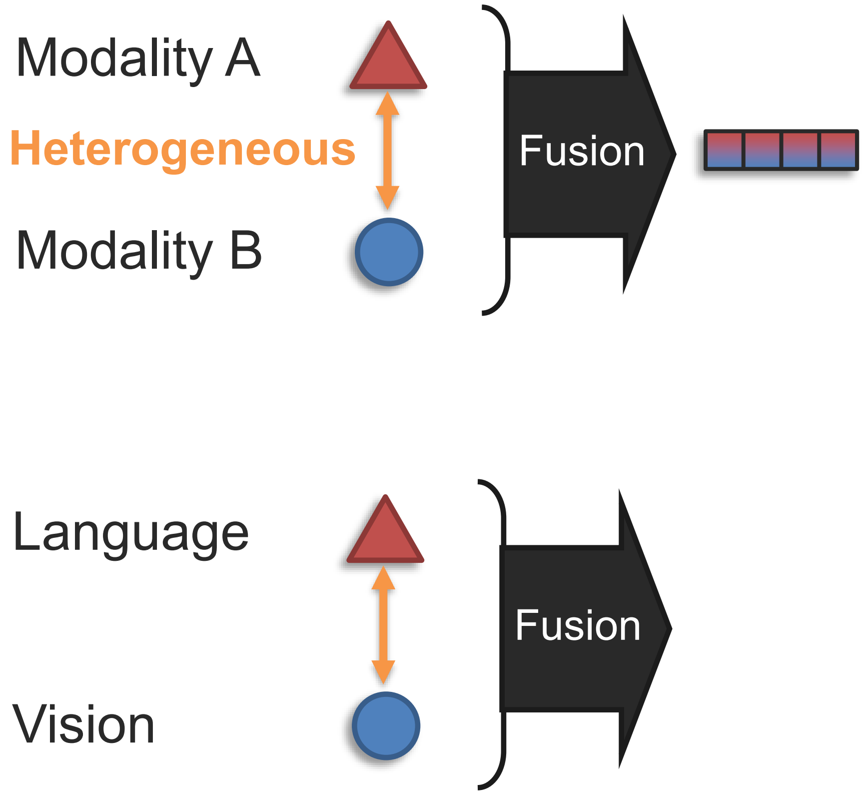
$$\mathcal{L}(y,\hat{y}_{uni}) + \mathcal{L}(y-\hat{y}_{uni},\hat{y}_{bi}) + \mathcal{L}(y-\hat{y}_{uni}-\hat{y}_{bi},\hat{y}_{tri})$$



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

Fusion with Heterogeneous Modalities

Goal: Fusion with raw modalities



Can the same fusion algorithm handle raw heterogeneous modalities?

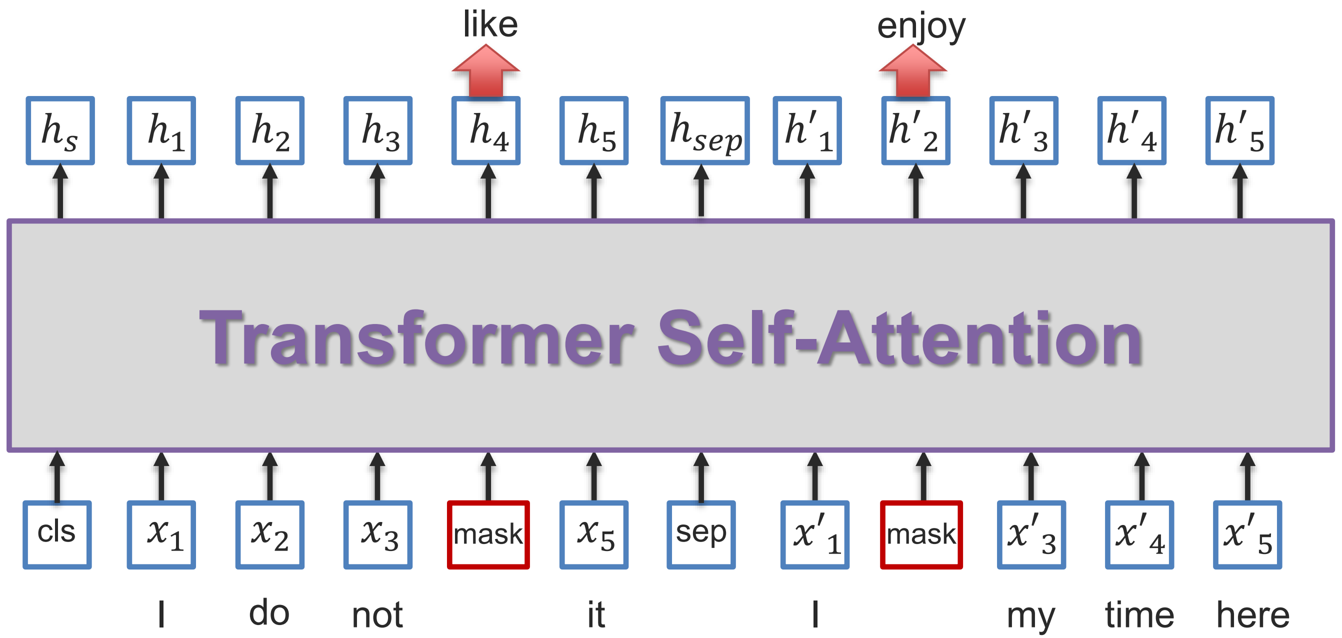
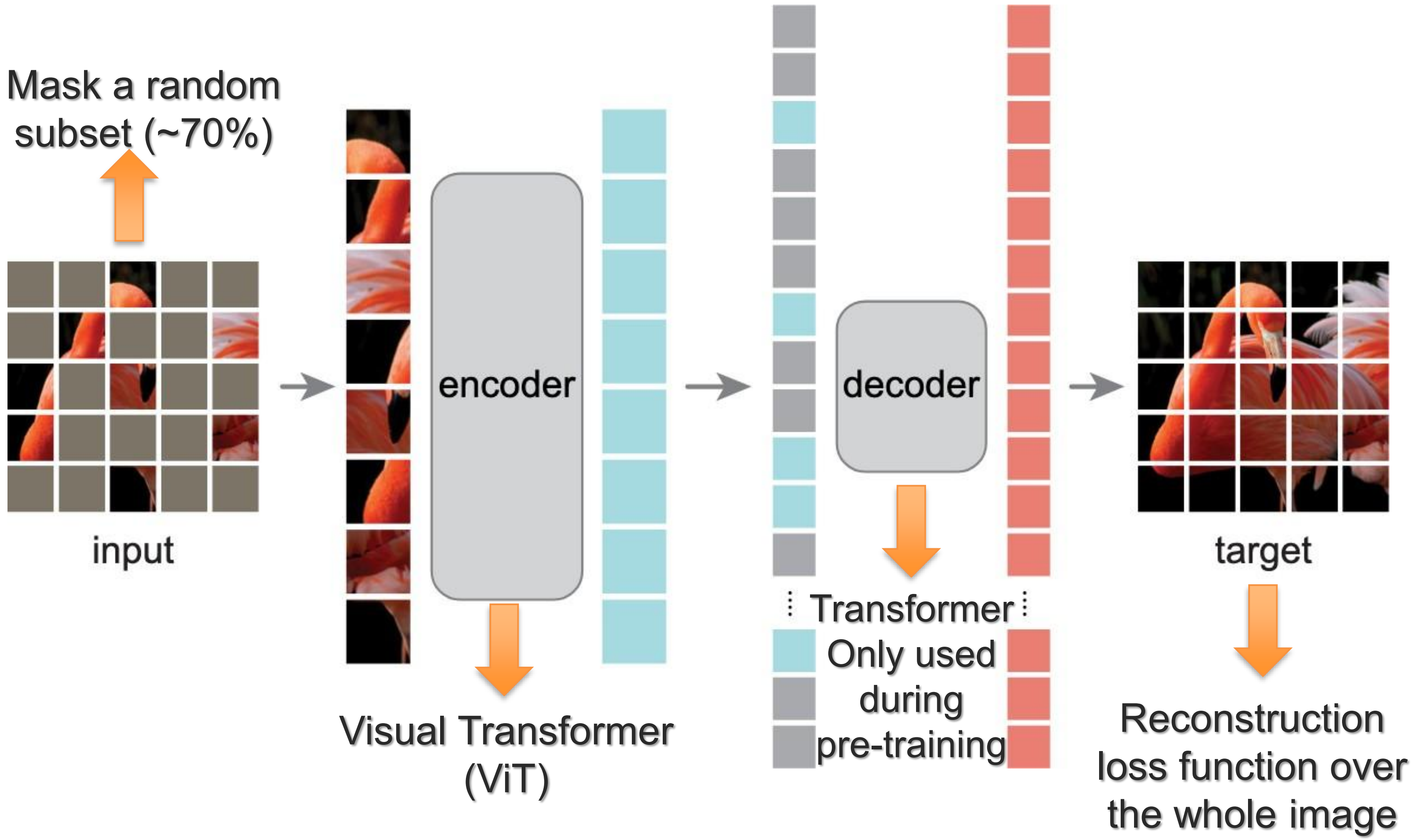
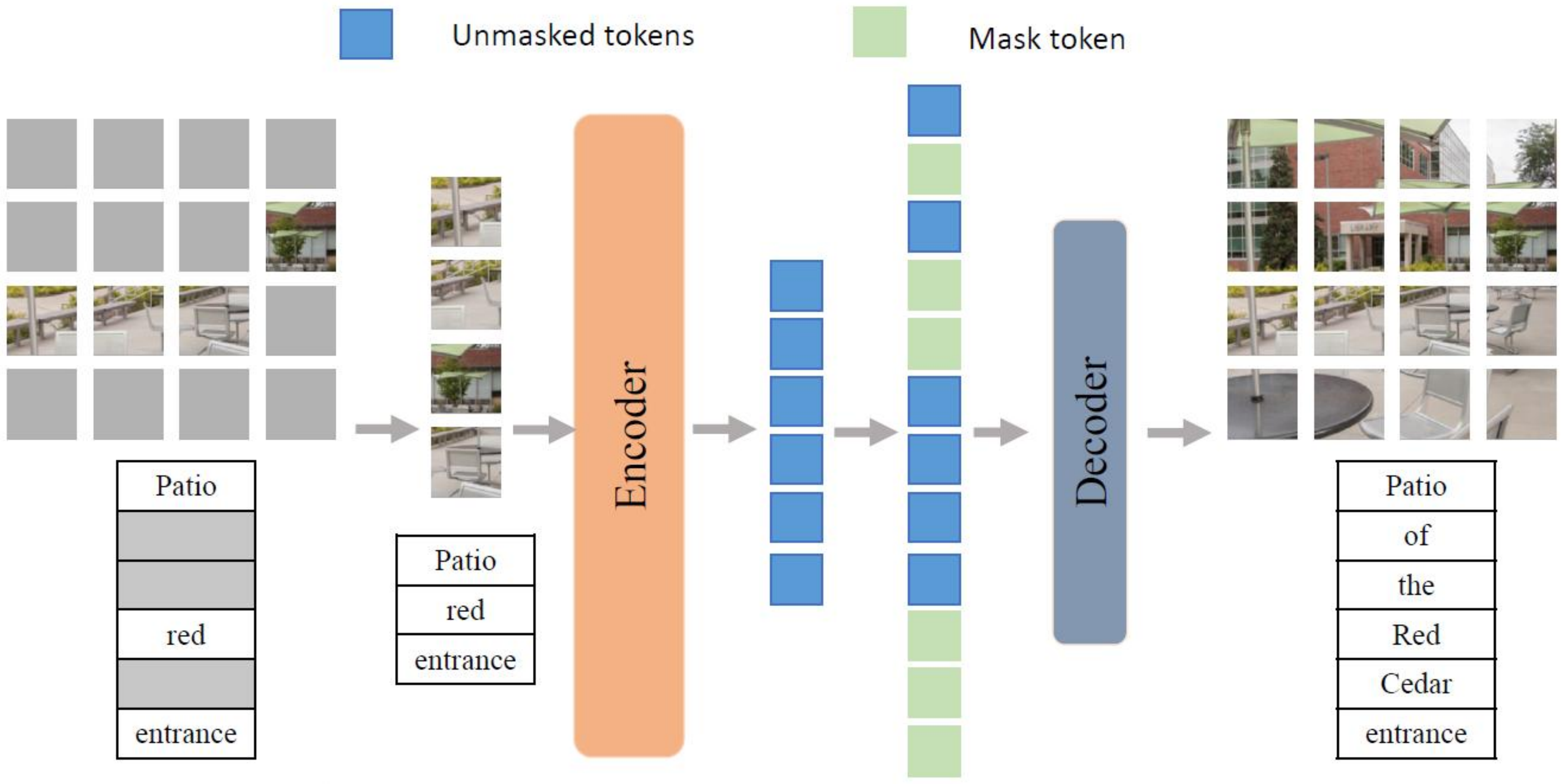


Image Representation Learning: Masked Auto-Encoder (MAE)



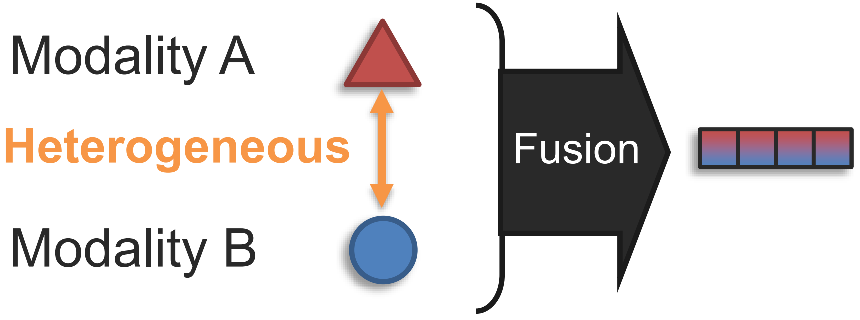
[He et al., Masked Autoencoders Are Scalable Vision Learners, CVPR 2022]

Multimodal Masked Autoencoder

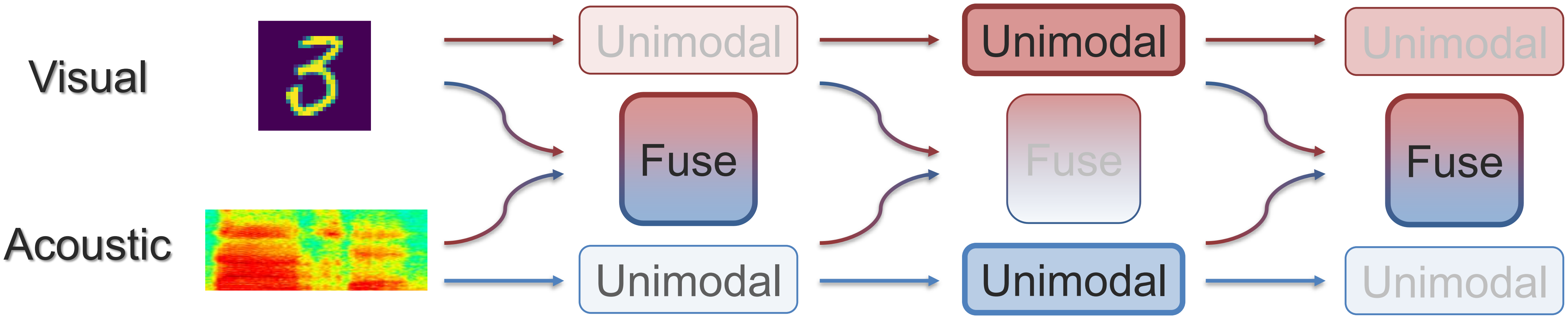


[Geng et al., Multimodal Masked Autoencoders Learn Transferable Representations, 2022]

Dynamic Early Fusion



Idea: Deciding when to fuse in early fusion



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

Dynamic Early Fusion

Fusion fully learned from optimization and data

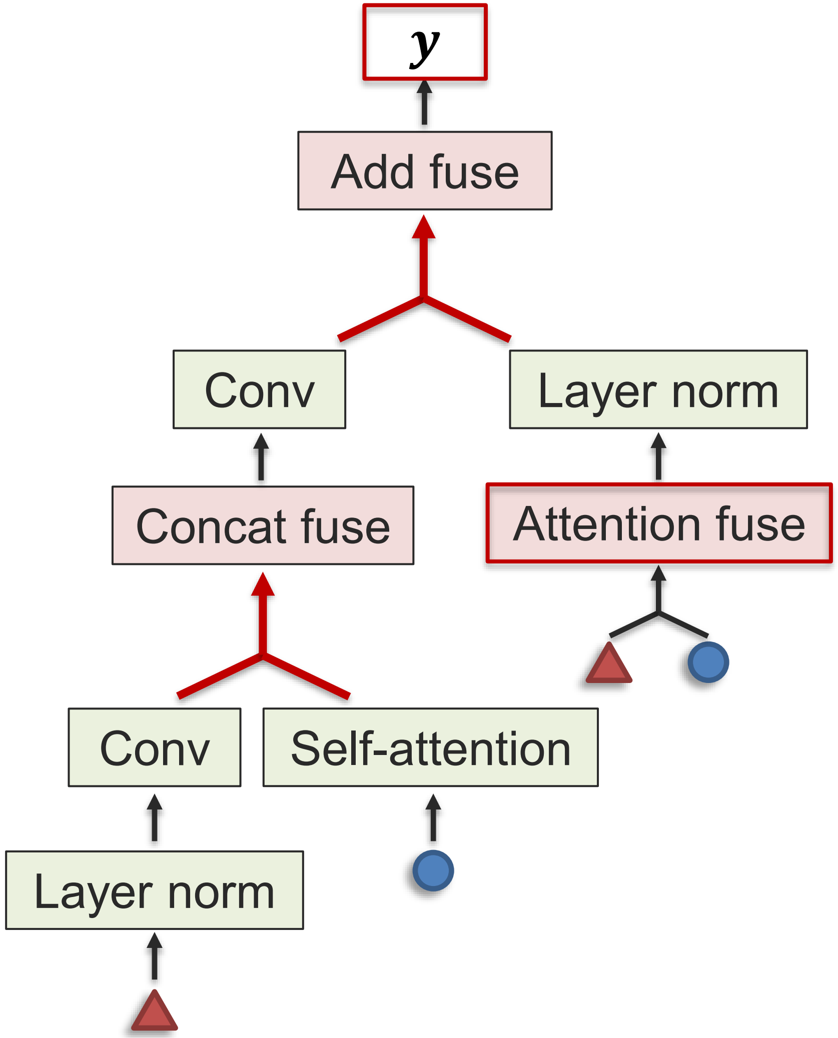
1. Define basic representation building blocks

ReLU | Layer norm | Conv | Self-attention

2. Define basic fusion building blocks

Concat fuse | Attention fuse | Add fuse

3. Automatically search for composition using neural architecture search

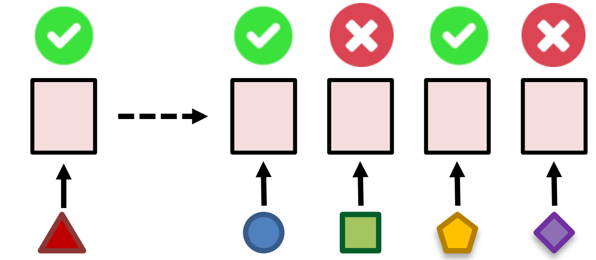


[Xu et al., MUFASA: Multimodal Fusion Architecture Search for Electronic Health Records. AAAI 2021]
[Liu et al., DARTS: Differentiable Architecture Search. ICLR 2019]

Heterogeneity-aware Fusion

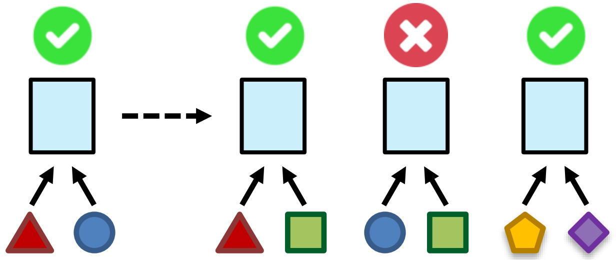
Information transfer, transfer learning perspective

1a. Estimate modality heterogeneity via transfer

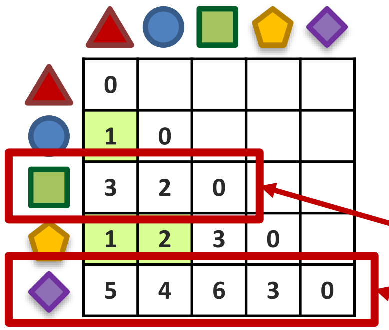


(Implicitly captures heterogeneity)

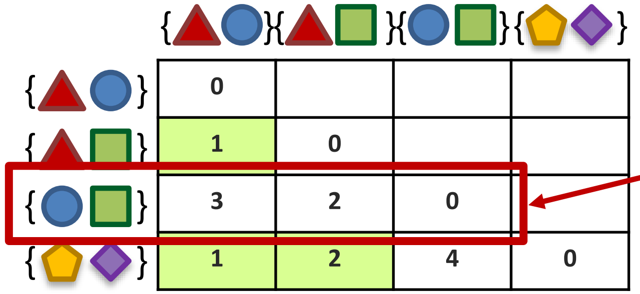
1b. Estimate interaction heterogeneity via transfer



2a. Compute modality heterogeneity matrix



2b. Compute interaction heterogeneity matrix



3. Determine parameter clustering

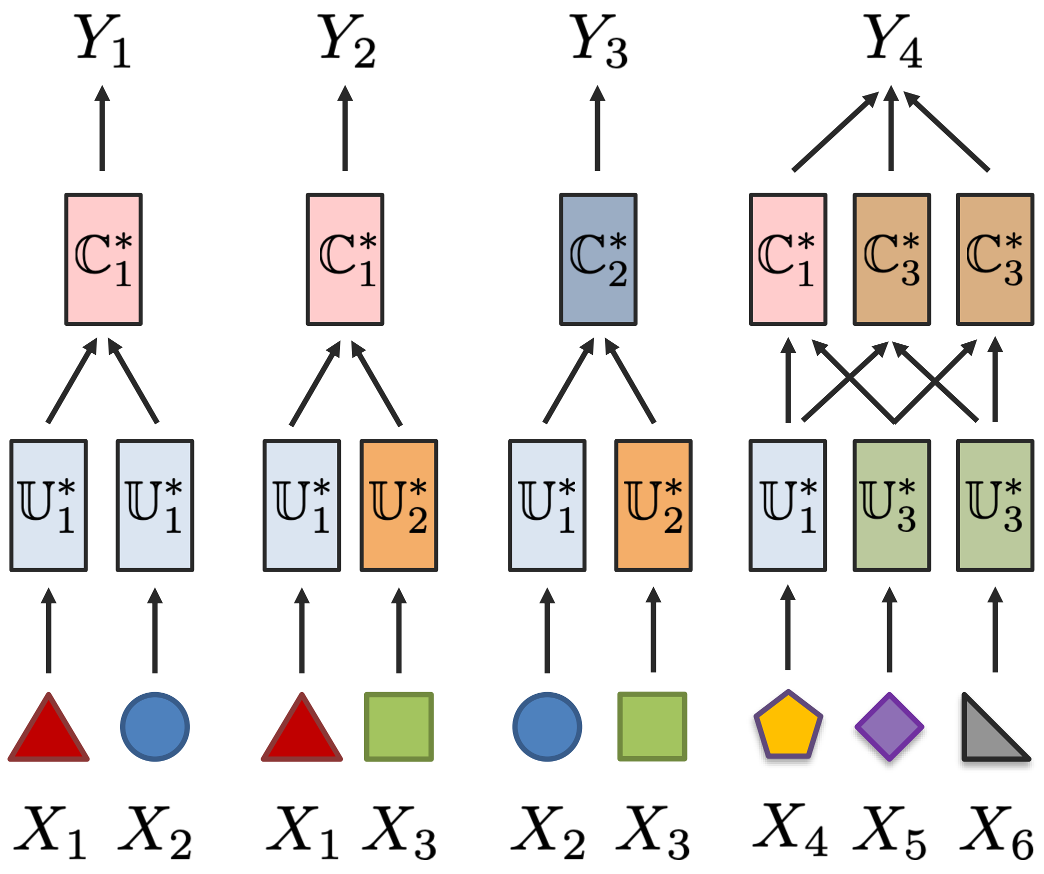
$$\mathbb{U}_1 = \{U_1, U_2, U_4\}$$
$$\mathbb{U}_2 = \{U_3\}$$
$$\mathbb{U}_3 = \{U_5\}$$

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

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

Heterogeneity-aware Fusion

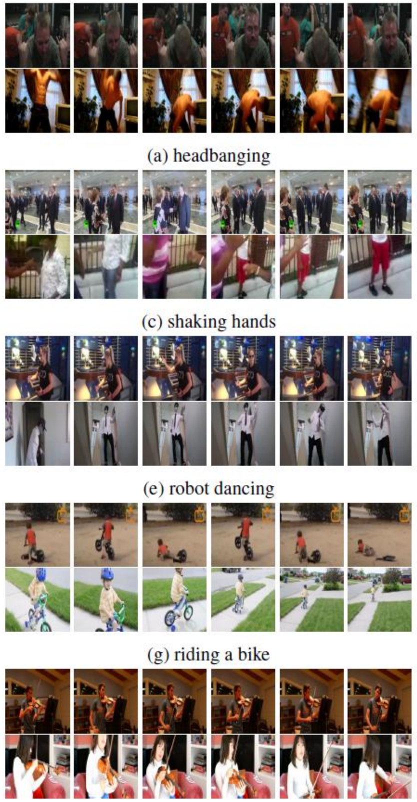
Information transfer, transfer learning perspective



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

Improving Optimization

Kinetics dataset



Adding more modalities should always help?

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

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

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

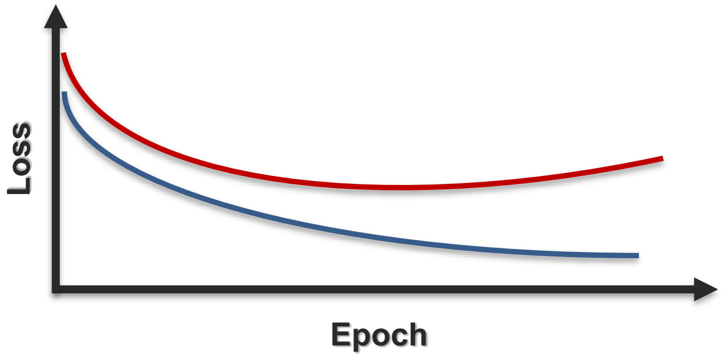
[Wang et al., What Makes Training Multi-modal Classification Networks Hard? CVPR 2020]
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]

Improving Optimization

Relevance heterogeneity

2 explanations for drop in performance:

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



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

Gap between training and valid loss

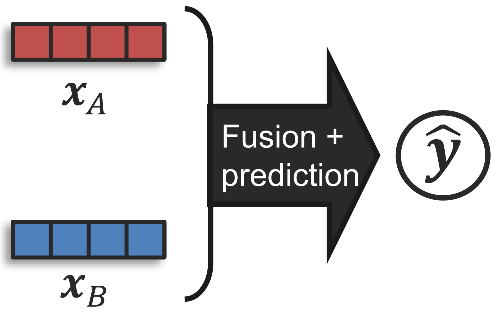
OGR wrt each modality tells us how much to train that modality

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

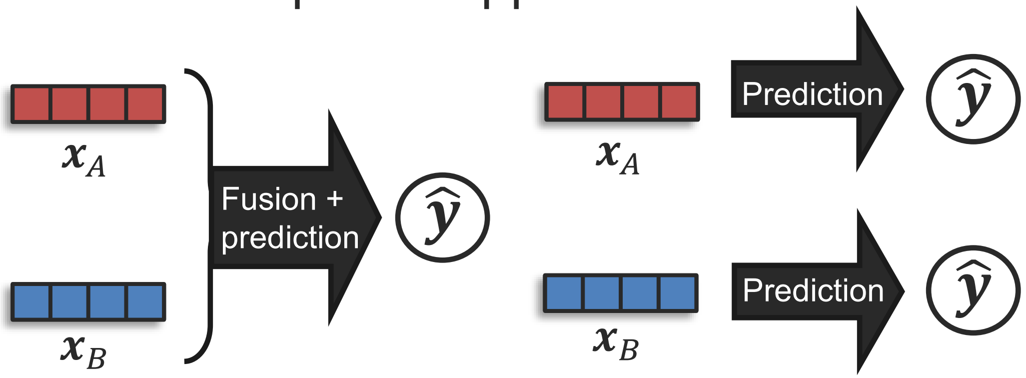
Improving Optimization

Relevance heterogeneity

Conventional approach



Proposed approach



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

Reweight multimodal loss using unimodal OGR values

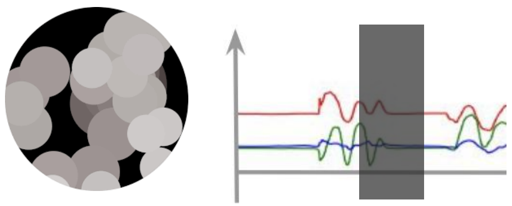
Allows to better balance generalization & overfitting rate of different modalities

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

Heterogeneity in Noise: Studying Robustness

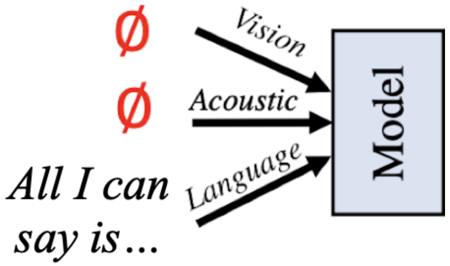
Noise within Modality

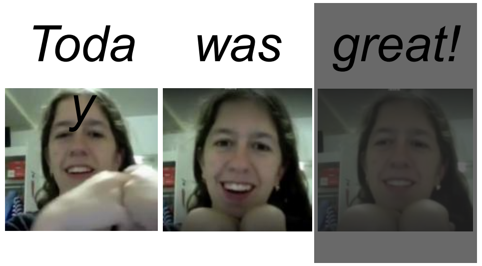
noise → nosie



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

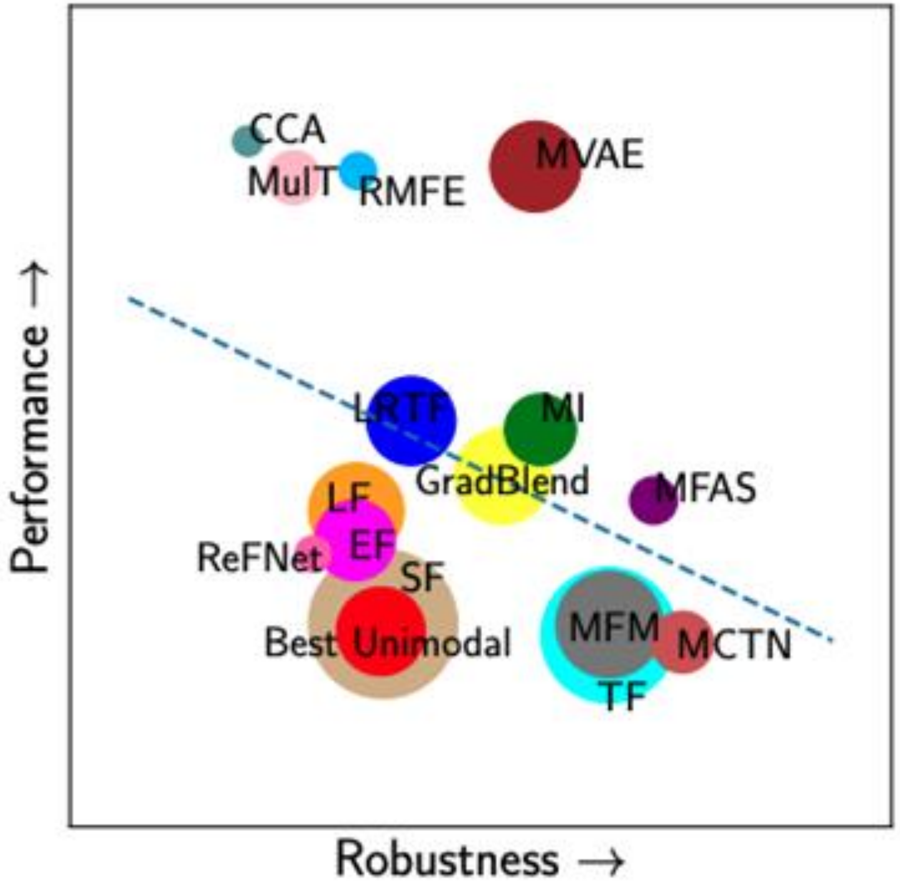
Missing Modalities





[Zadeh et al., 2020]

Strong tradeoffs between performance and robustness



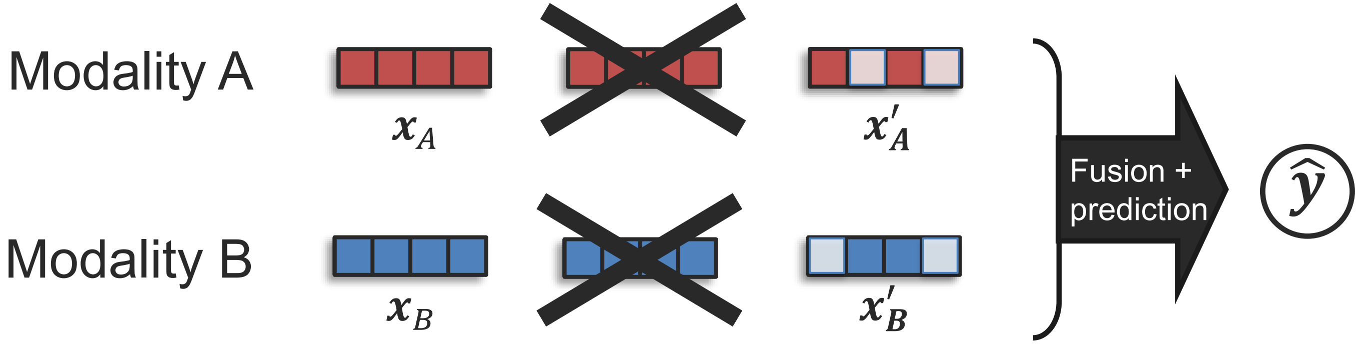
rate of accuracy drops

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

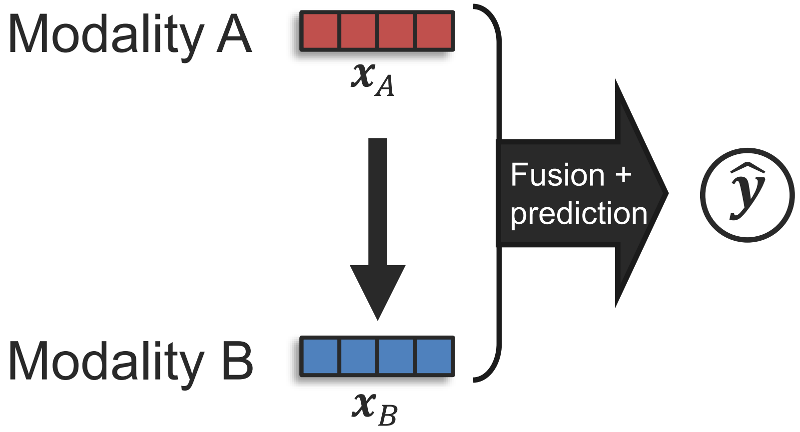
Heterogeneity in Noise: Studying Robustness

Several approaches towards more robust models

Robust data + training



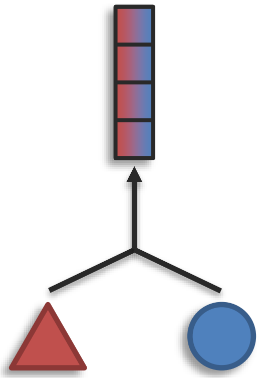
Infer missing modalities



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 by Cyclic Translations Between Modalities. AAAI 2019

Sub-Challenge 1a: Representation Fusion



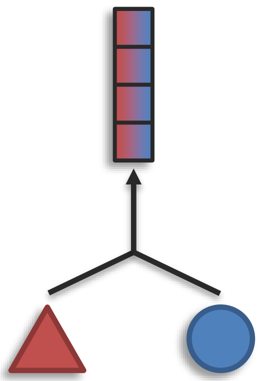
Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Challenge 1: Representation

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

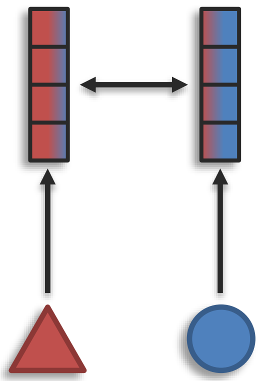
Sub-challenges:

Fusion



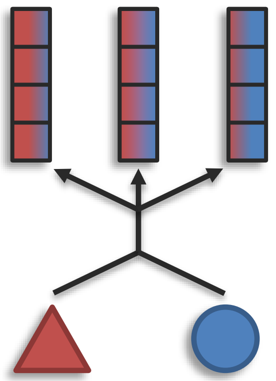
modalities > # representations

Coordination



modalities = # representations

Fission



modalities < # representations

Next week