Multimodal Learning

IFT6758 - Data Science

Sources:

Slides are mostly from CMU Multimodal Communication and Machine Learning Laboratory [MultiCompLab]

https://towardsdatascience.com/ensemble-methods-bagging-boosting-and-stacking-c9214a10a205

https://www.coursehero.com/file/22624681/771A-lec21-slides/

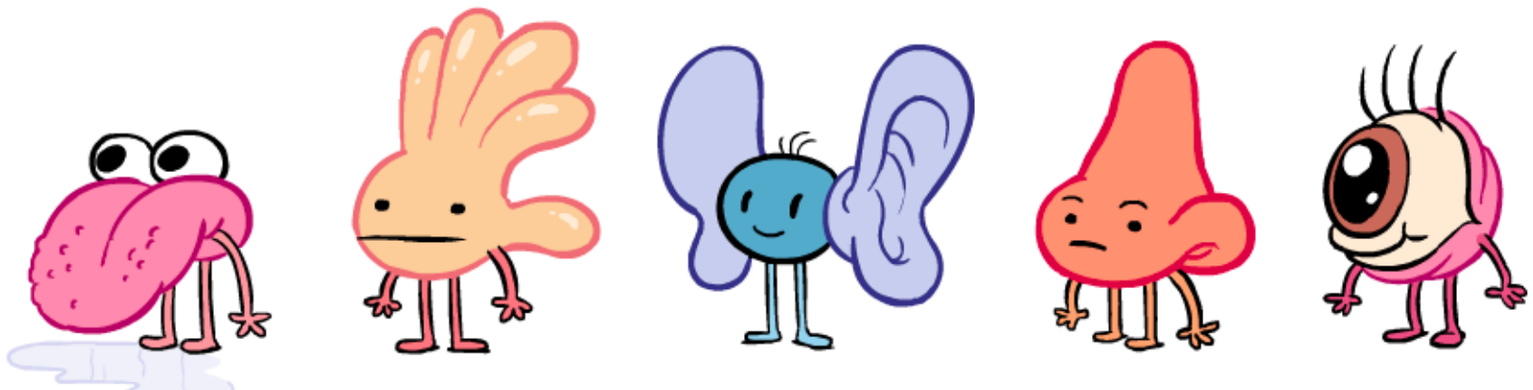
https://arxiv.org/abs/1705.09406

https://www.cs.princeton.edu/courses/archive/spring16/cos495/

Mila

Université de Montréal

What is Multimodal Learning?



- “Our experience of the world is multimodal - we see objects, hear sounds, feel texture, smell odors, and taste flavors.”

- “In order for Artificial Intelligence to make progress in understanding the world around us, it needs to be able to interpret such multimodal signals together. Multimodal machine learning aims to build models that can process and relate information from multiple modalities.”

What is Multimodal?

Modality

The way in which something happens or is experienced.

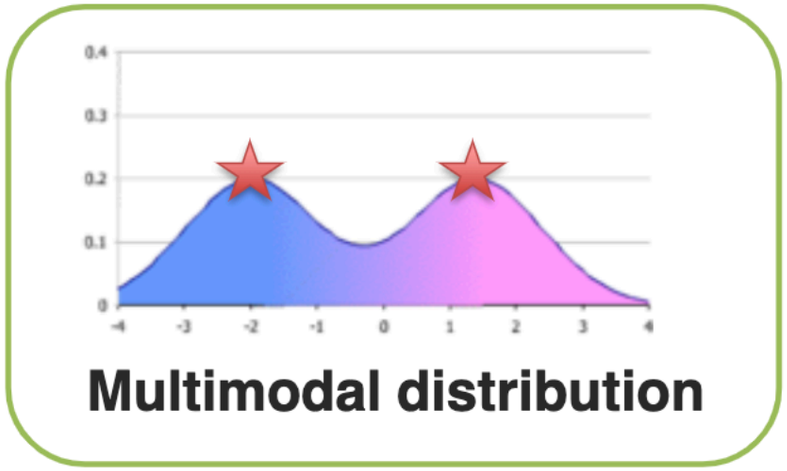
- Modality refers to a certain type of information and/or the representation format in which information is stored.
- Sensory modality: one of the primary forms of sensation, as vision or touch; channel of communication.

Medium (“middle”)

A means or instrumentality for storing or communicating information; system of communication/transmission.

- Medium is the means whereby this information is delivered to the senses of the interpreter.

What is Multimodal?



- Multiple modes, i.e., distinct “peaks” (local maxima) in the probability density function

Multimodal Communicative Behaviors

Verbal

Lexicon
- Words

Syntax
- Part-of-speech
- Dependencies

Pragmatics
- Discourse acts

Vocal

Prosody
- Intonation
- Voice quality

Vocal expressions
- Laughter, moans

Visual

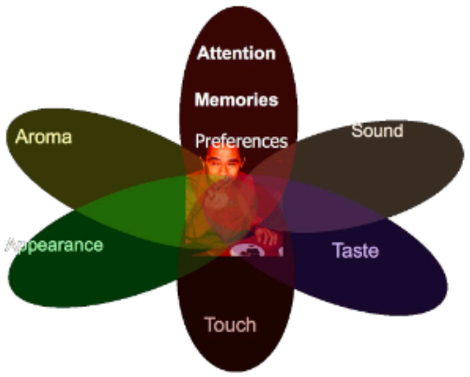
Gestures
- Head gestures
- Eye gestures
- Arm gestures

Body language
- Body posture
- Proxemics

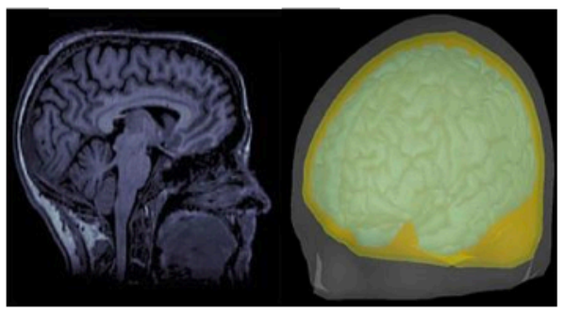
Eye contact
- Head gaze
- Eye gaze

Facial expressions
- FACS action units
- Smile, frowning

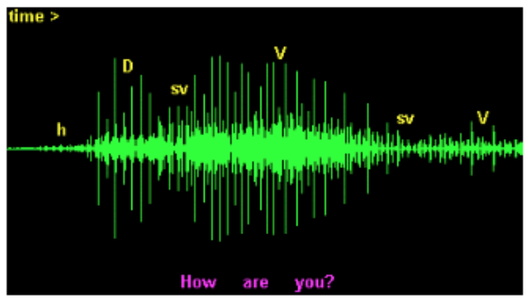
Multiple modalities



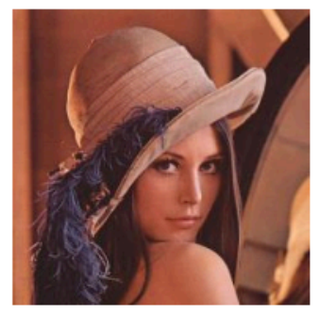
Psychology



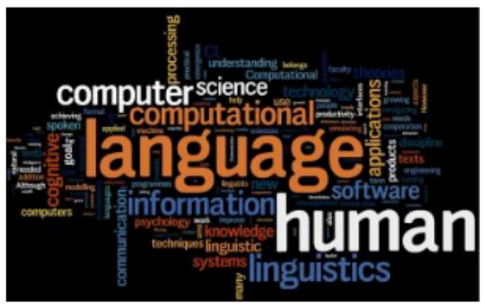
Medical



Speech

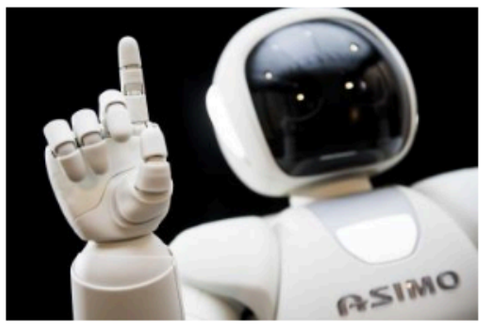


Vision

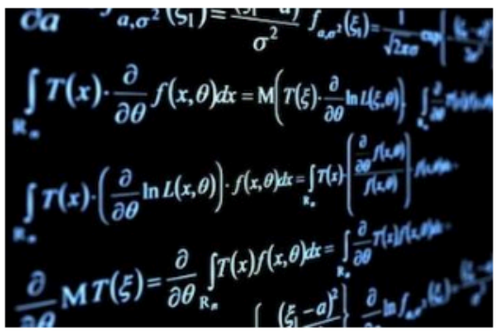


Language

Multimedia



Robotics



Learning

Examples of Modalities

Natural language (both spoken or written)

Visual (from images or videos)

Auditory (including voice, sounds and music)

Haptics / touch

Smell, taste and self-motion

Physiological signals

- Electrocardiogram (ECG), skin conductance

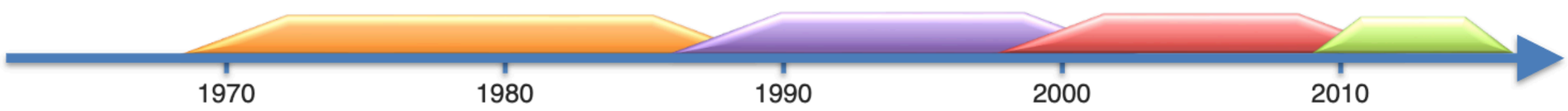
Other modalities

- Infrared images, depth images, fMRI

Prior research on Multimodal

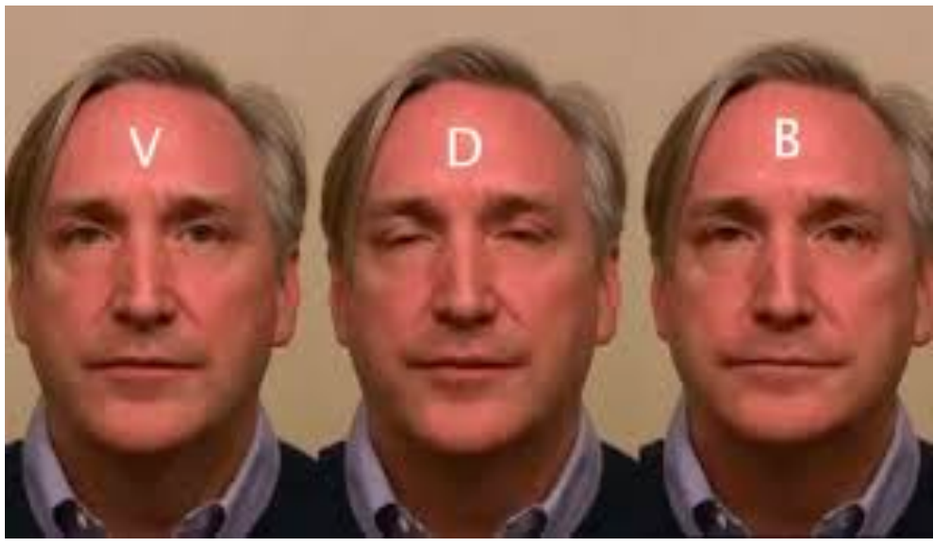
Four eras of multimodal research

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

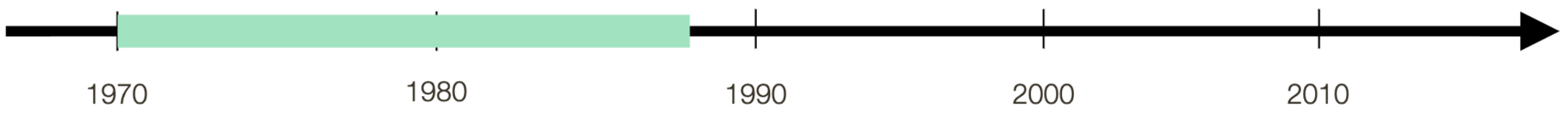


McGurk Effect

The McGurk Effect



McGurk Effect (1976)



Multimodal real-world tasks

- Affect recognition
  - Emotion
  - Persuasion
  - Personality traits
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

Core technichal challanges

- **Representation**
- **Alignment**
- **Fusion**
- **Translation**
- **Co-learning**

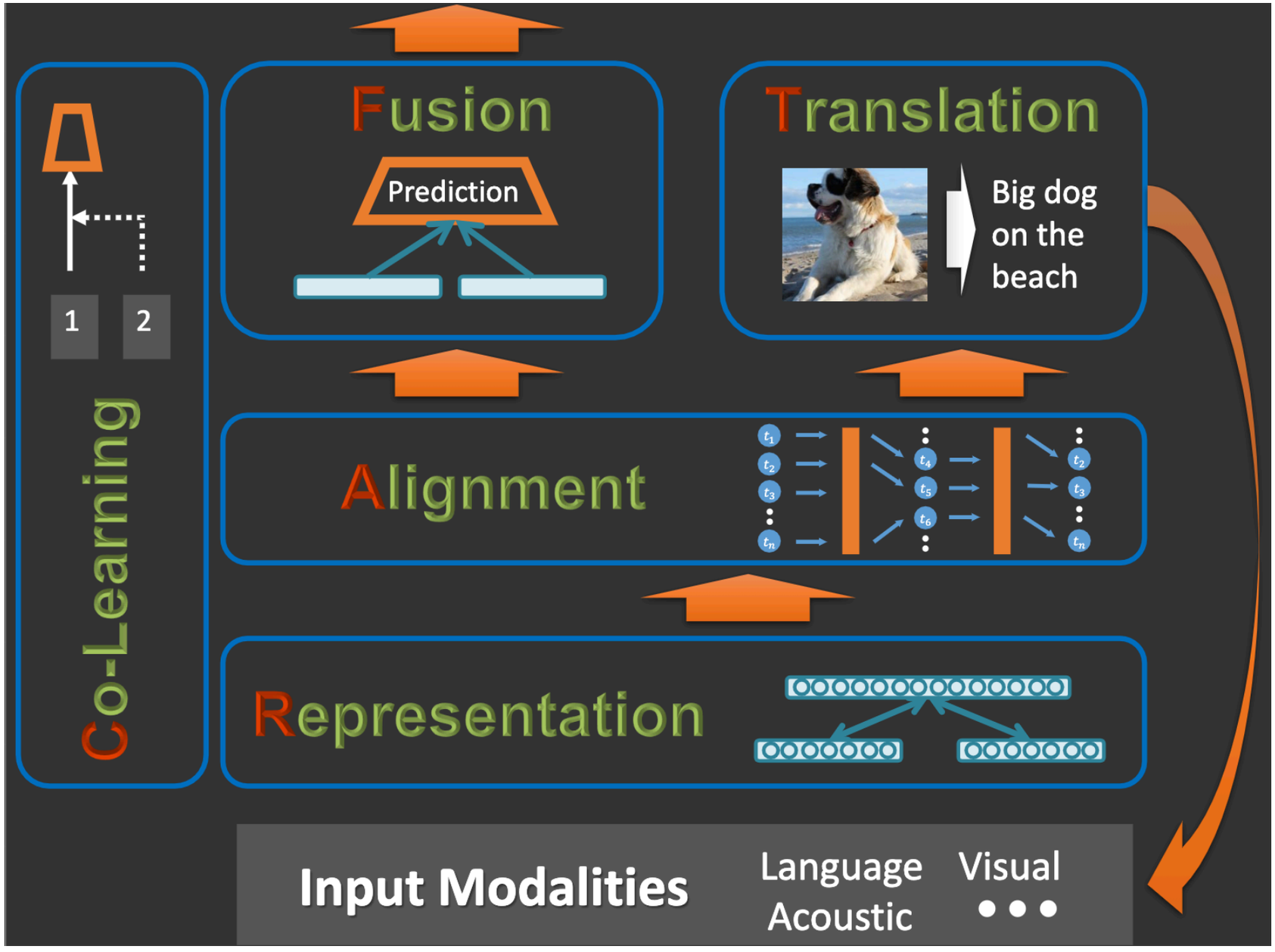
Multimodal Machine Learning: A Survey and Taxonomy

By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

https://arxiv.org/abs/1705.09406

☑ 5 core challenges
☑ 37 taxonomic classes
☑ 253 referenced citations

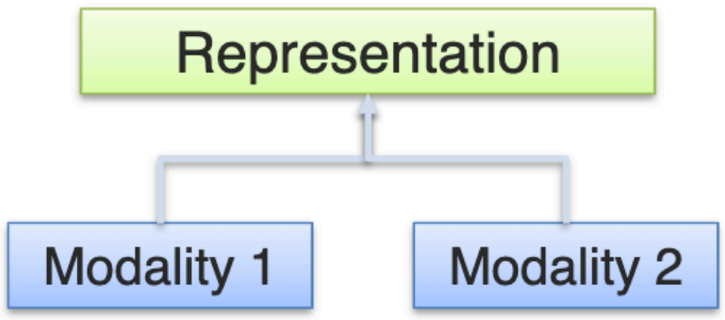
Architecture



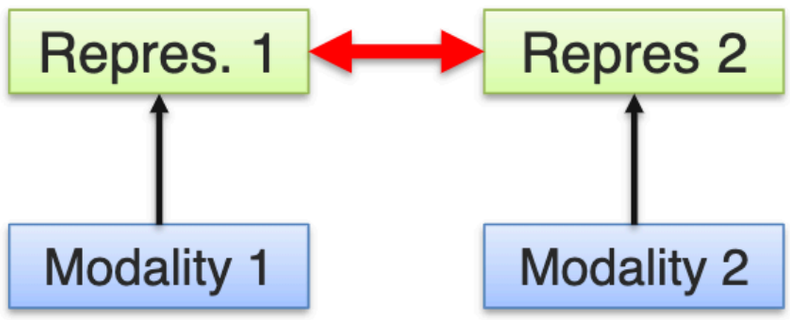
Core Challenge 1: Multimodal representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

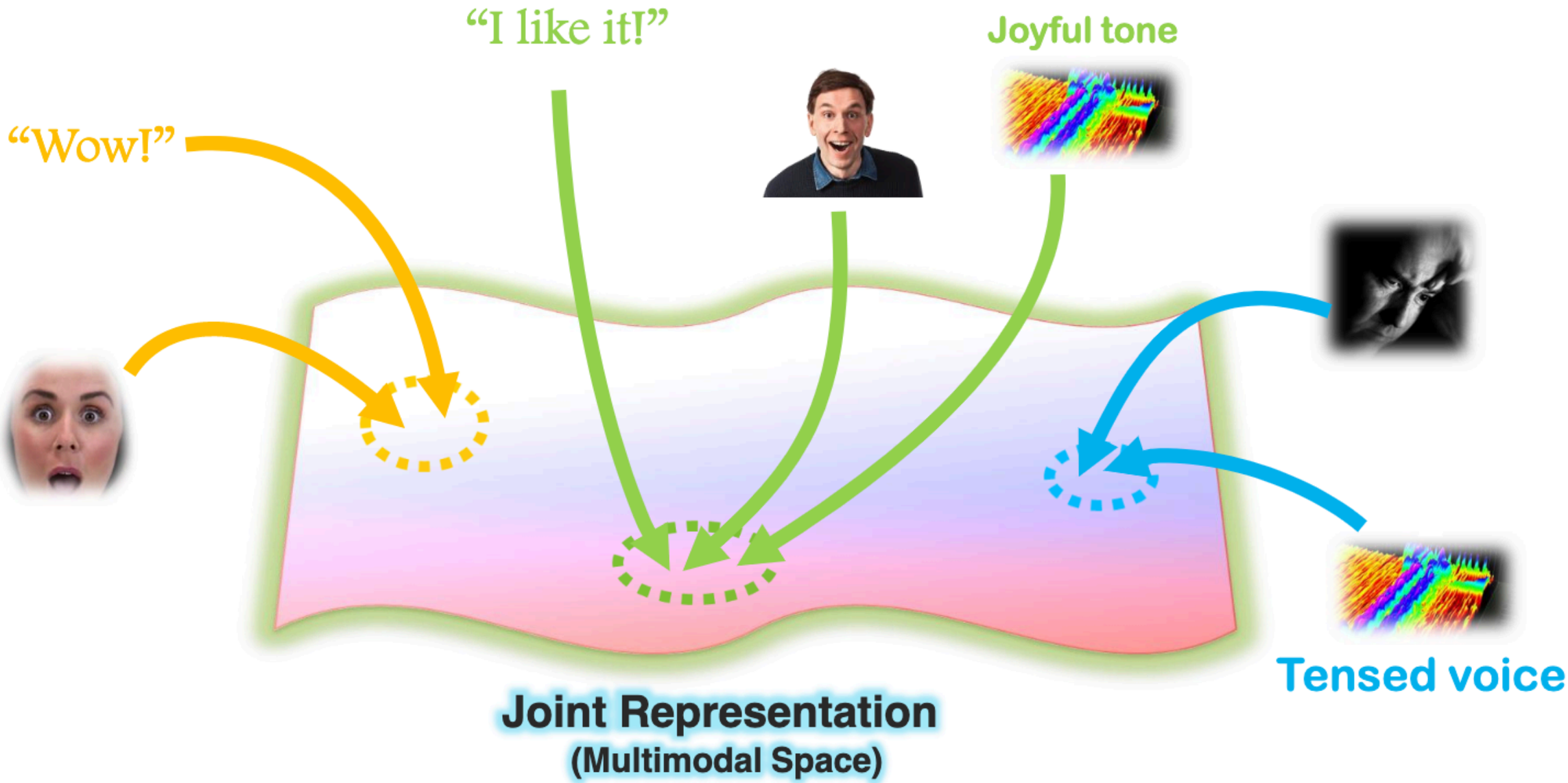
A Joint representations:



B Coordinated representations:



Joint Multimodal Representation



Joint Multimodal Representation

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

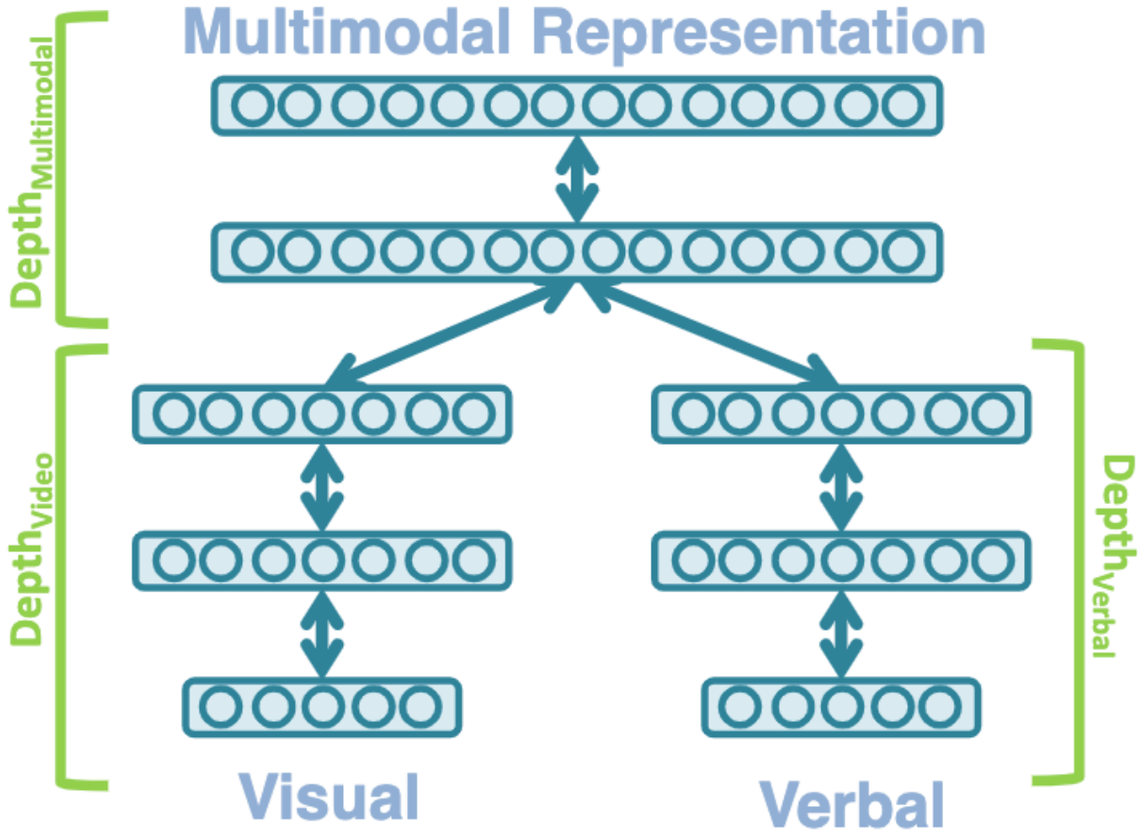
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

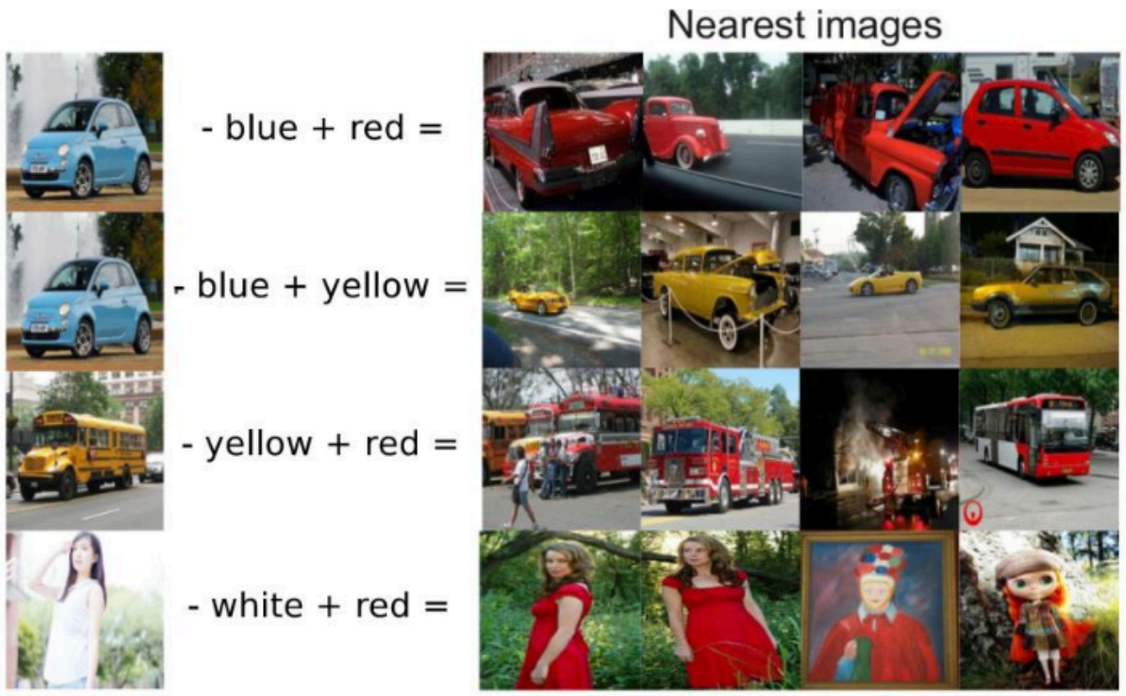
Audio-visual emotion recognition

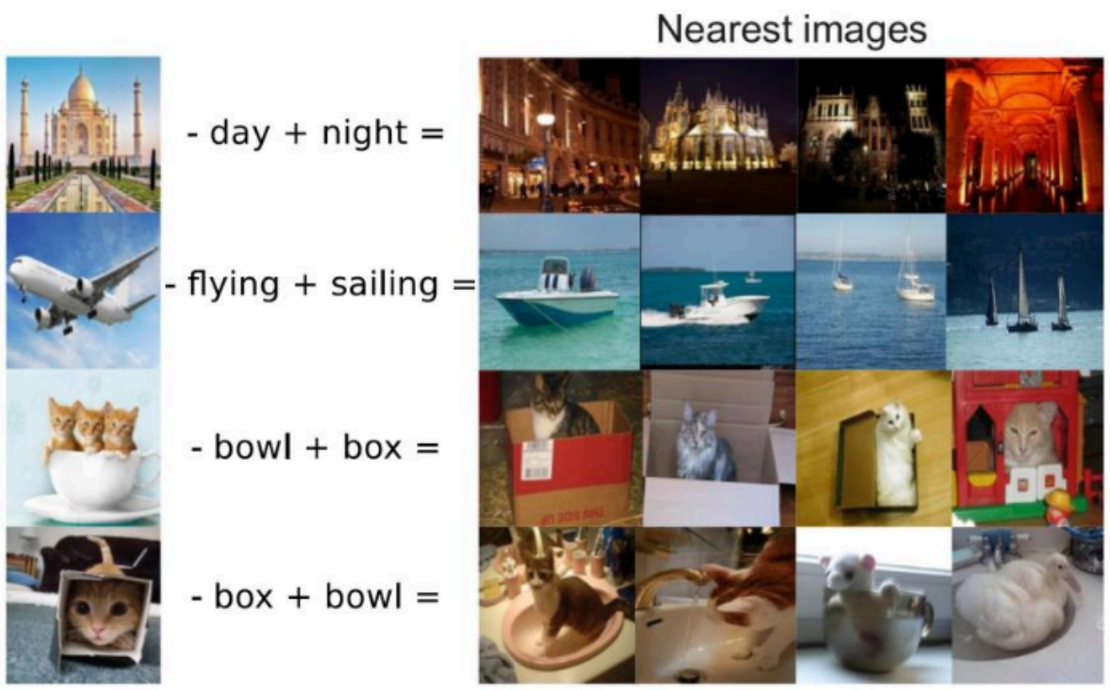
[Kim et al., ICASSP 2013]

- Deep Boltzmann Machine



Multimodal Vector Space Arithmetic

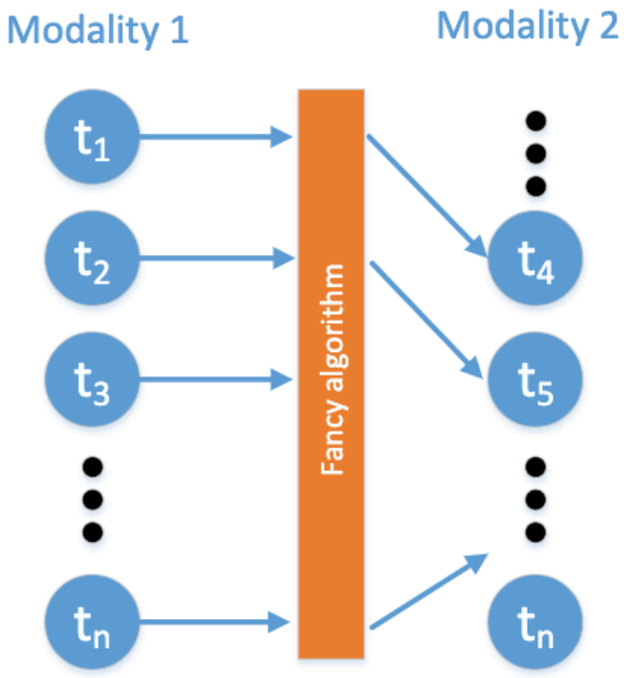




[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Core Challenge 2: Alignment

Definition: Identify the direct relations between (sub)elements from two or more different modalities.



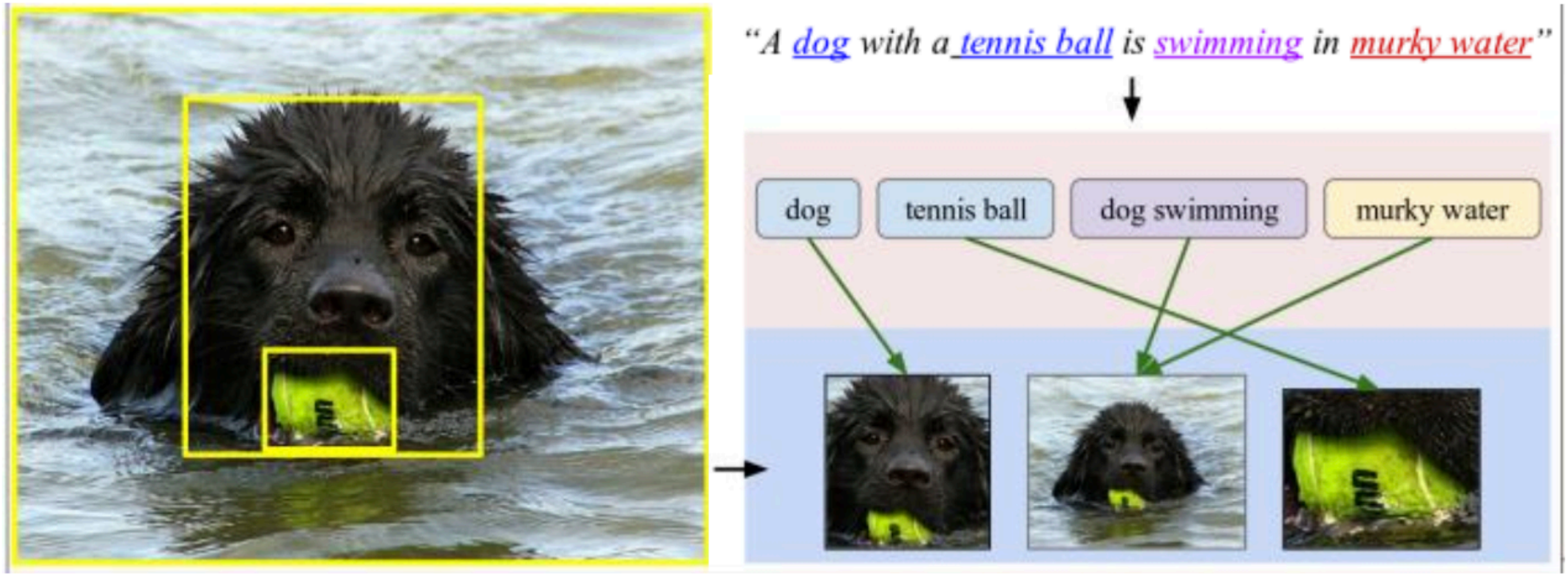
A Explicit Alignment

The goal is to directly find correspondences between elements of different modalities

B Implicit Alignment

Uses internally latent alignment of modalities in order to better solve a different problem

Alignment



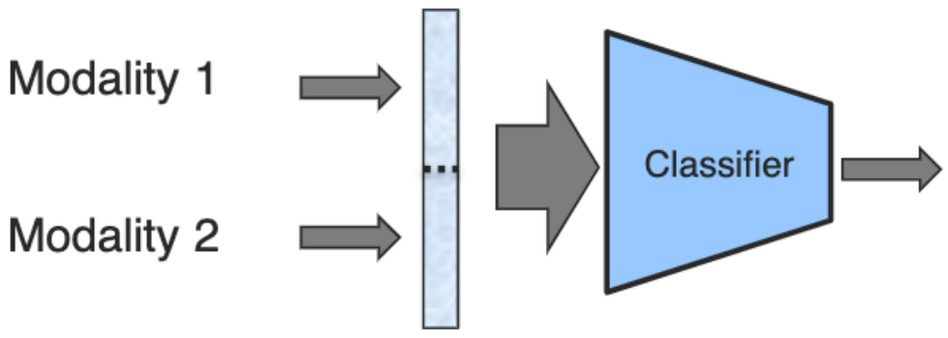
Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Core Challenge 3: Fusion

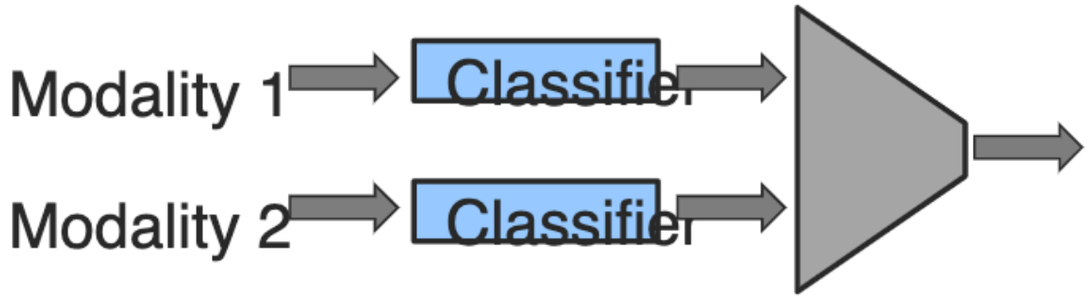
Definition: To join information from two or more modalities to perform a prediction task.

A Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion



Core Challenge 3: Fusion

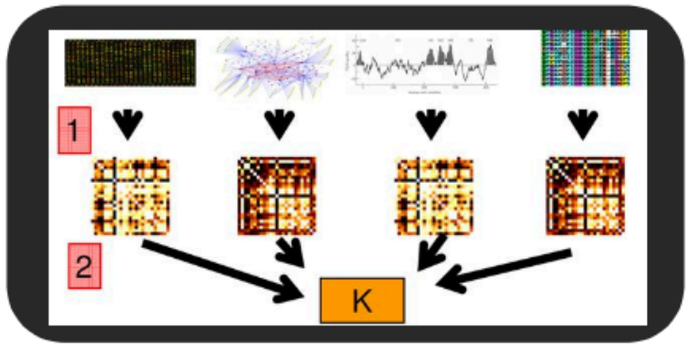
Definition: To join information from two or more modalities to perform a prediction task.

B Model-Based (Intermediate) Approaches

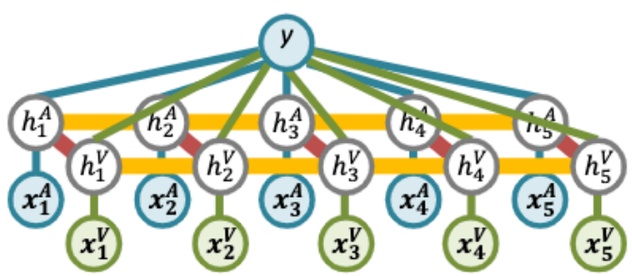
1) Deep neural networks

2) Kernel-based methods

3) Graphical models



Multiple kernel learning

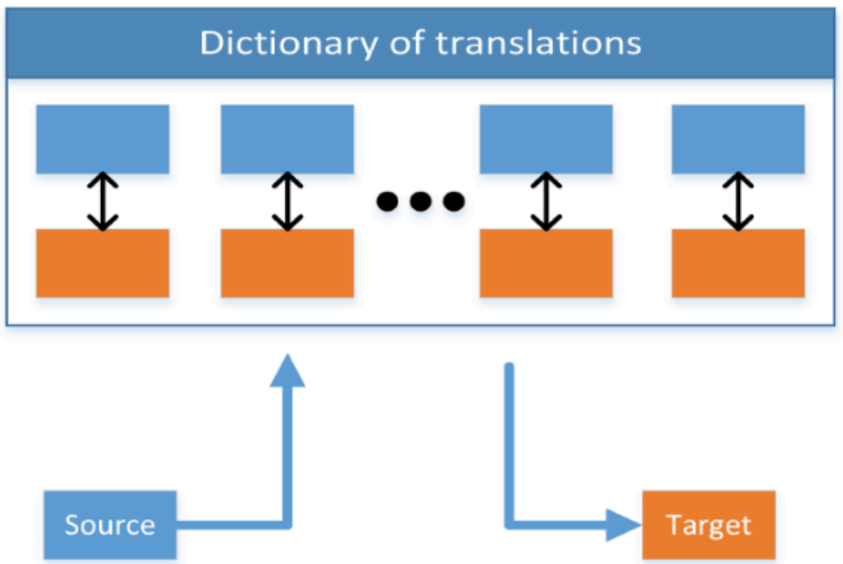


Multi-View Hidden CRF

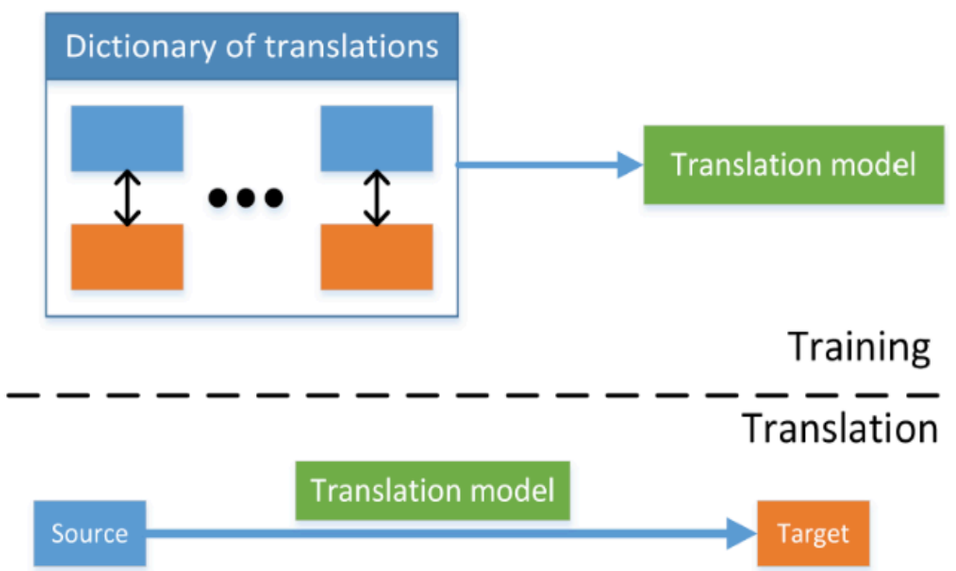
Core Challenge 4: Translation

Definition: Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.

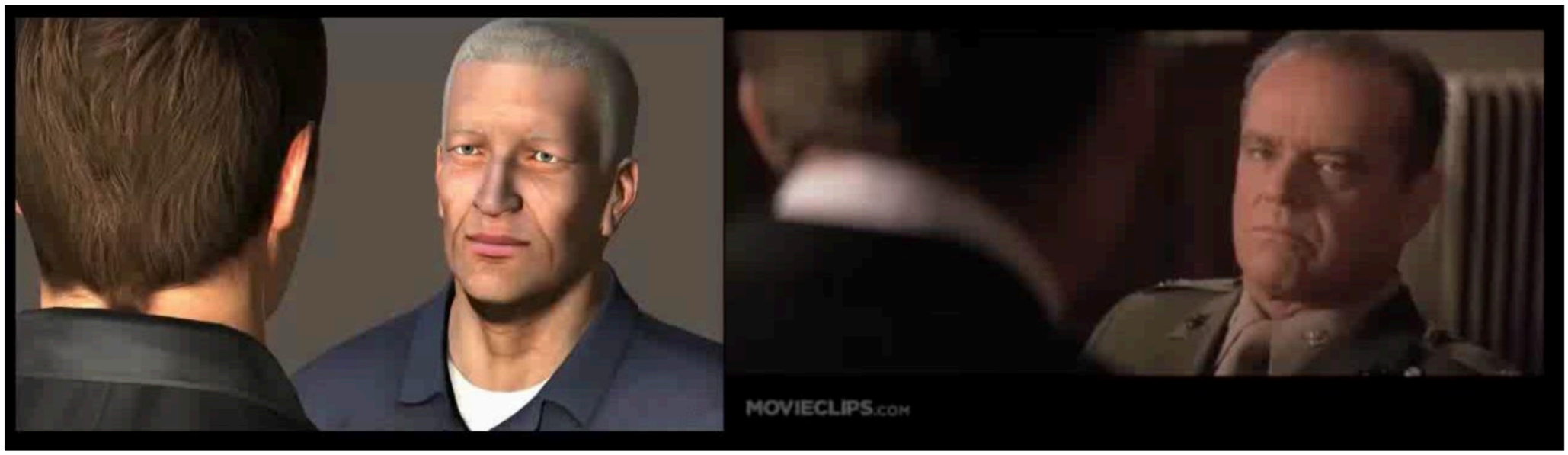
A Example-based



B Model-driven



Core Challenge 4: Translation

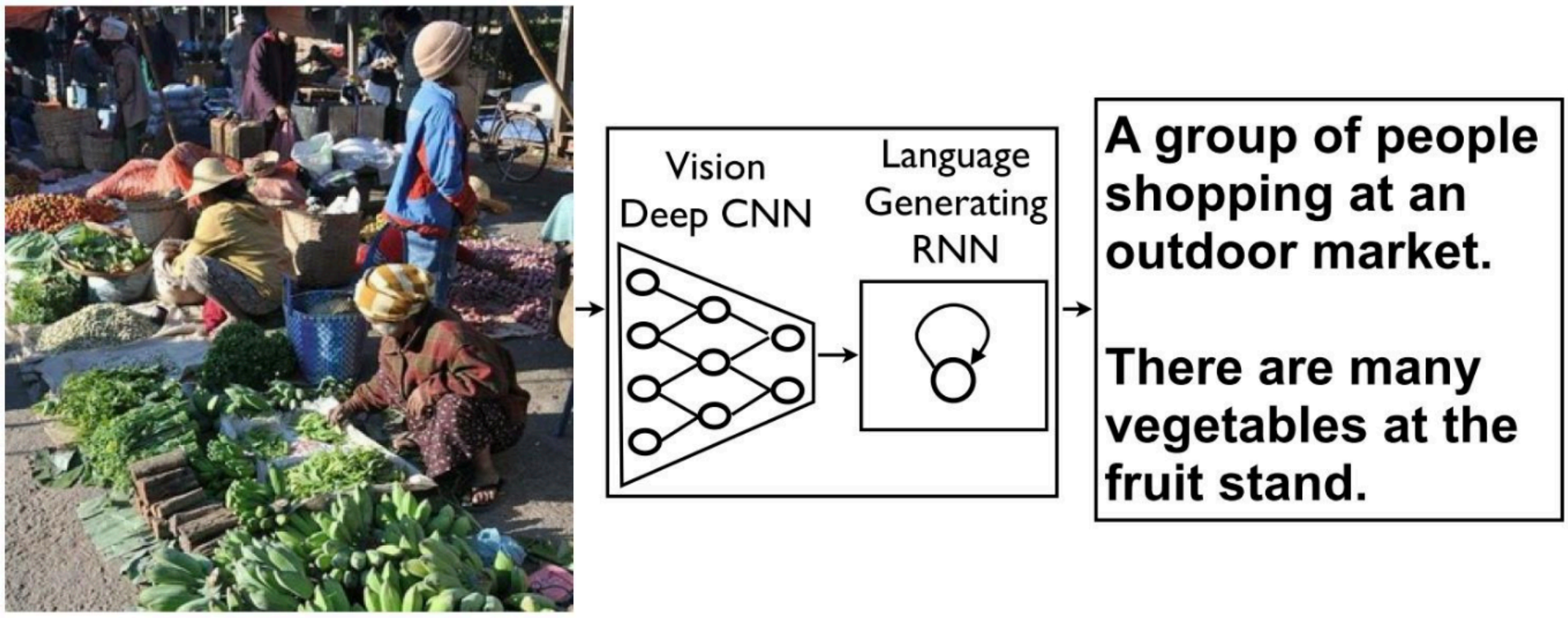


Visual gestures
(both speaker and listener gestures)

Transcriptions
+
Audio streams

Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

Core Challenge 4: Translation



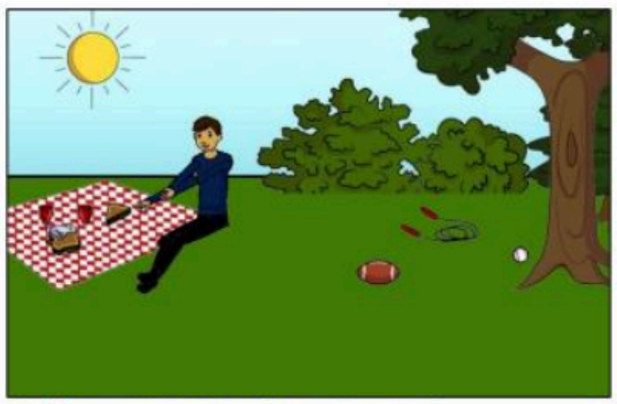
[Vinyals et al., “Show and Tell: A Neural Image Caption Generator”, CVPR 2015]

Core Challenge 4:
Translation (Visual Question Answering)

- A very new and exciting task created in part to address evaluation problems with the above task
- Task - Given an image and a question answer the question (http://www.visualqa.org/)

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

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

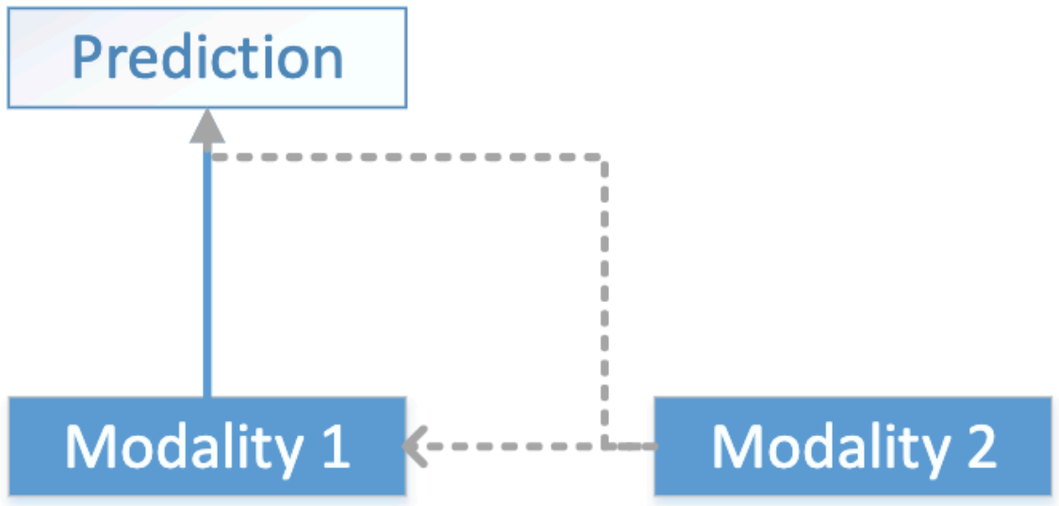


Is this person expecting company?
What is just under the tree?

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

Core Challenge 5: Co-Learning

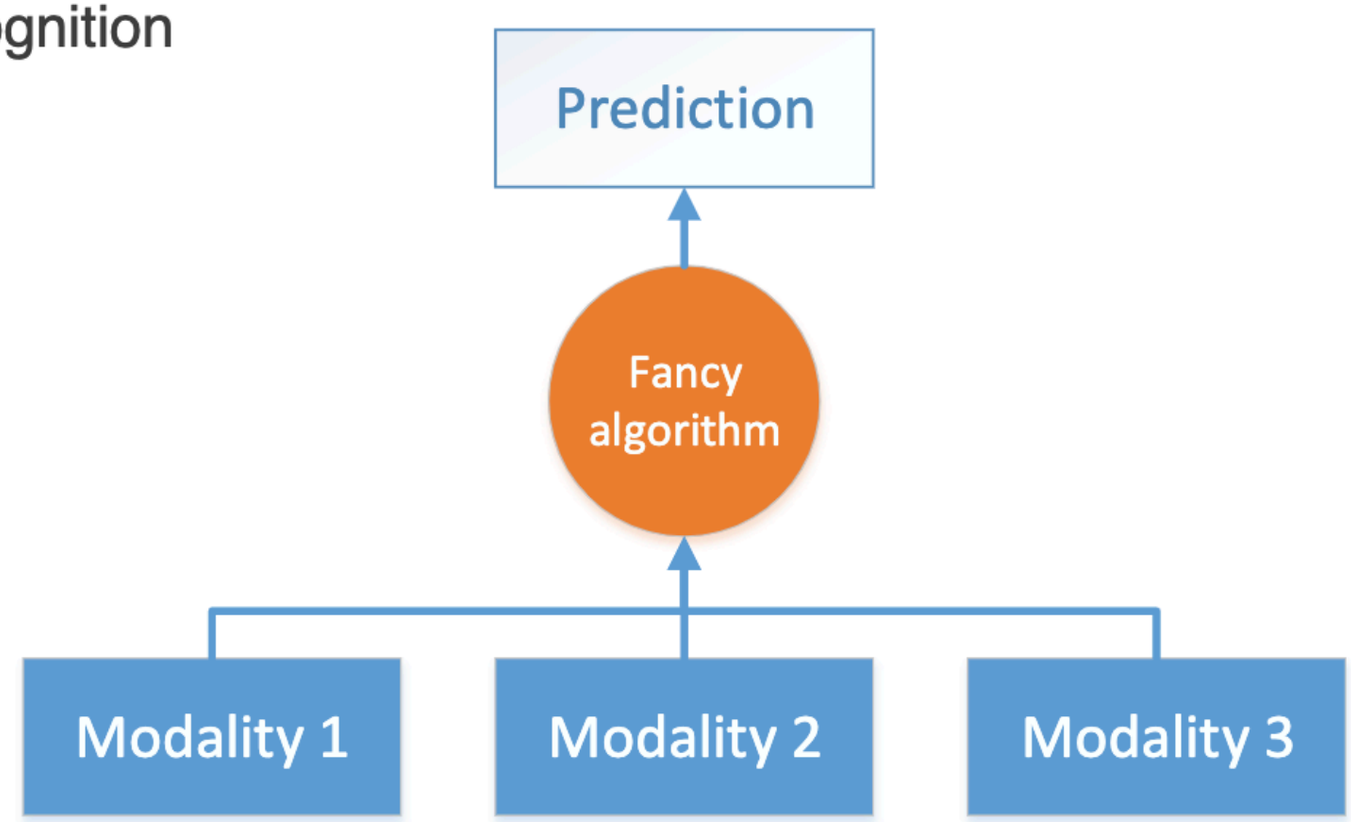
Definition: Transfer knowledge between modalities, including their representations and predictive models.



Multimodal Fusion

Mutimodal Fusion

- Process of joining information from two or more modalities to perform a prediction
  - One of the earlier and more established problems
  - e.g. audio-visual speech recognition, multimedia event detection, multimodal emotion recognition
- Two major types
- Model Free
  - Early, late, hybrid
- Model Based
  - Kernel Methods
  - Graphical models
  - Neural networks



Benefits

- Complementary information - McGurk effect
  - The sum is greater than the parts
- Robustness in presence of noise in one modality
- Dealing with missing or unobserved data in one of the modalities

- Examples
  - Audio-visual speech recognition
  - Audio-visual emotion recognition
  - Multimodal biometrics
  - Speaker identification and diarization

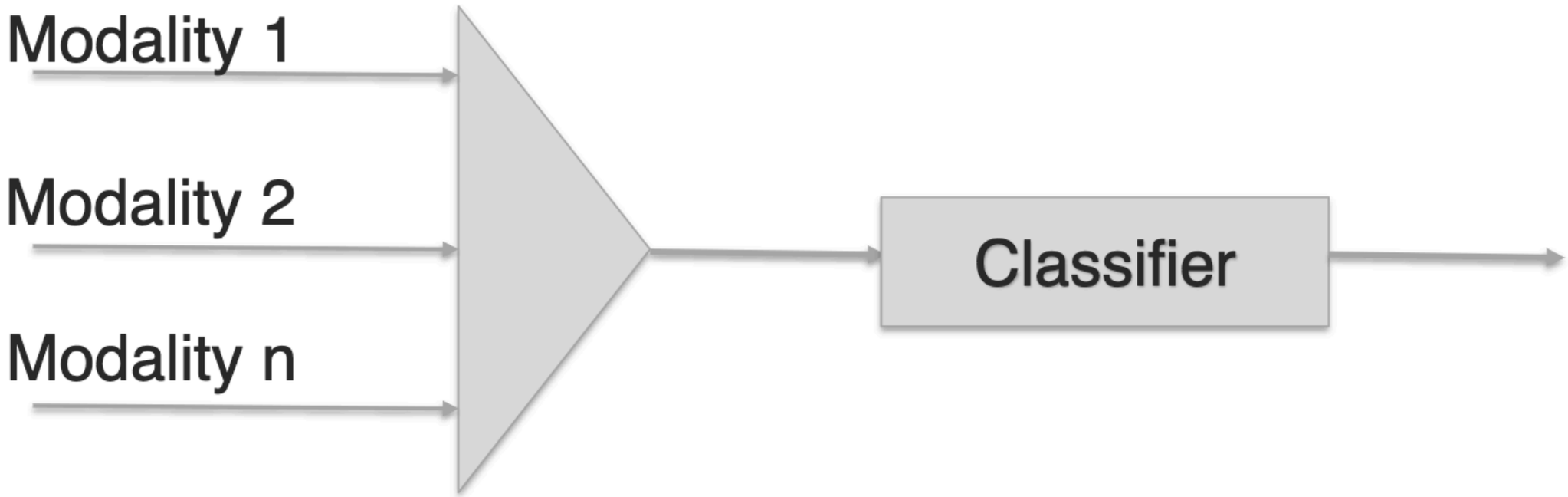
(a) answer-phone (a) get-out-car (a) fight-person

Challenges and pitfalls

- Different sampling rates
- Different amounts of noise in modalities
- Potentially missing data in one modality
- One of the modalities not being informative
- Different predictive power of each modality
- Modalities only providing redundancy

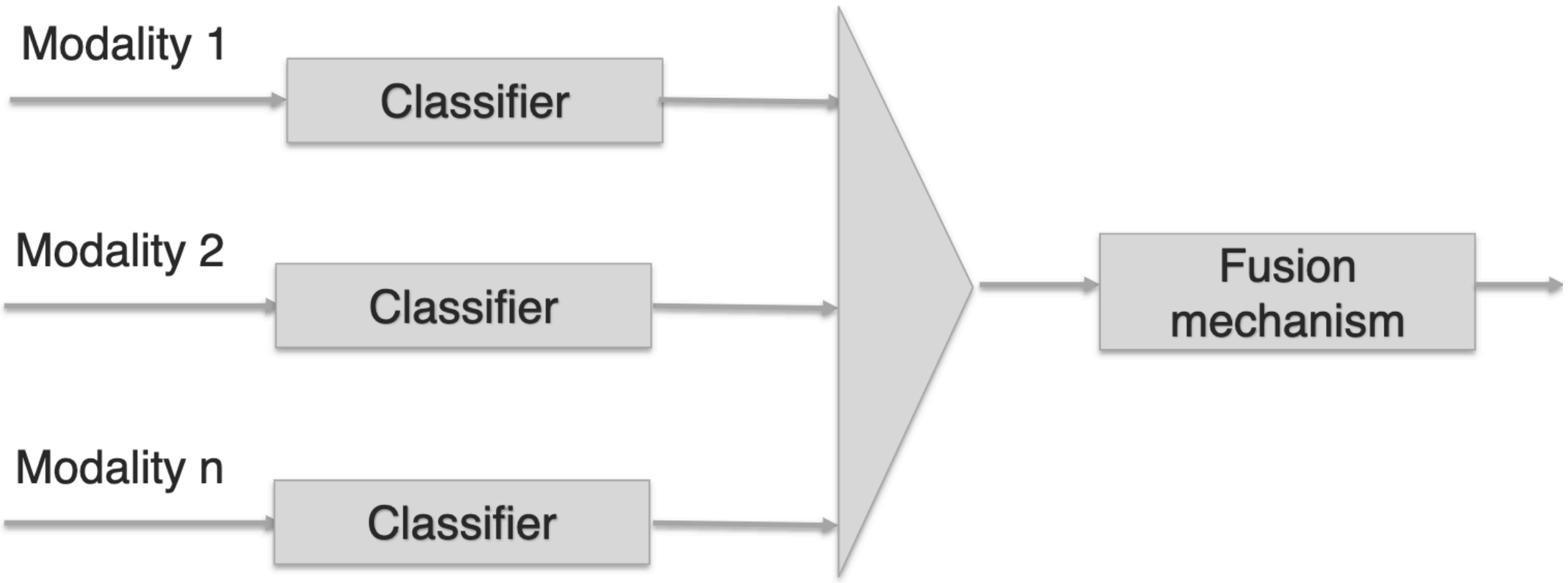
Model-free Fusion Approaches

Model-free approaches: Early Fusion



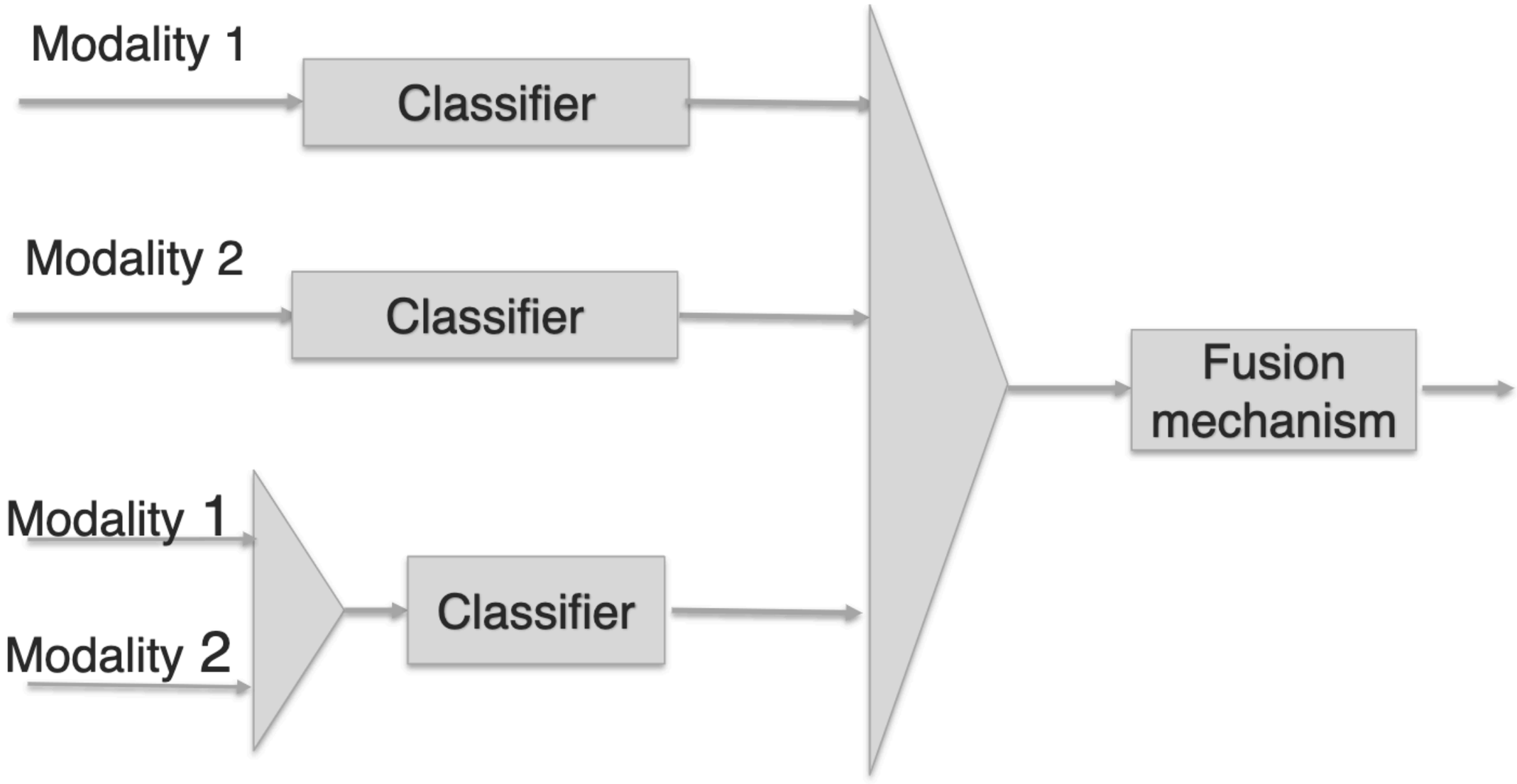
- Easy to implement – just concatenate the features
- Exploit dependencies between features
- Can end up very high dimensional
- More difficult to use if features have different framerates

Model-free approaches: Late Fusion



- Train a unimodal predictor and a multimodal fusion one
- Requires multiple training stages
- Do not model low level interactions between modalities
- Fusion mechanism can be voting, weighted sum or an ML approach

Model-free approaches: Hybrid Fusion

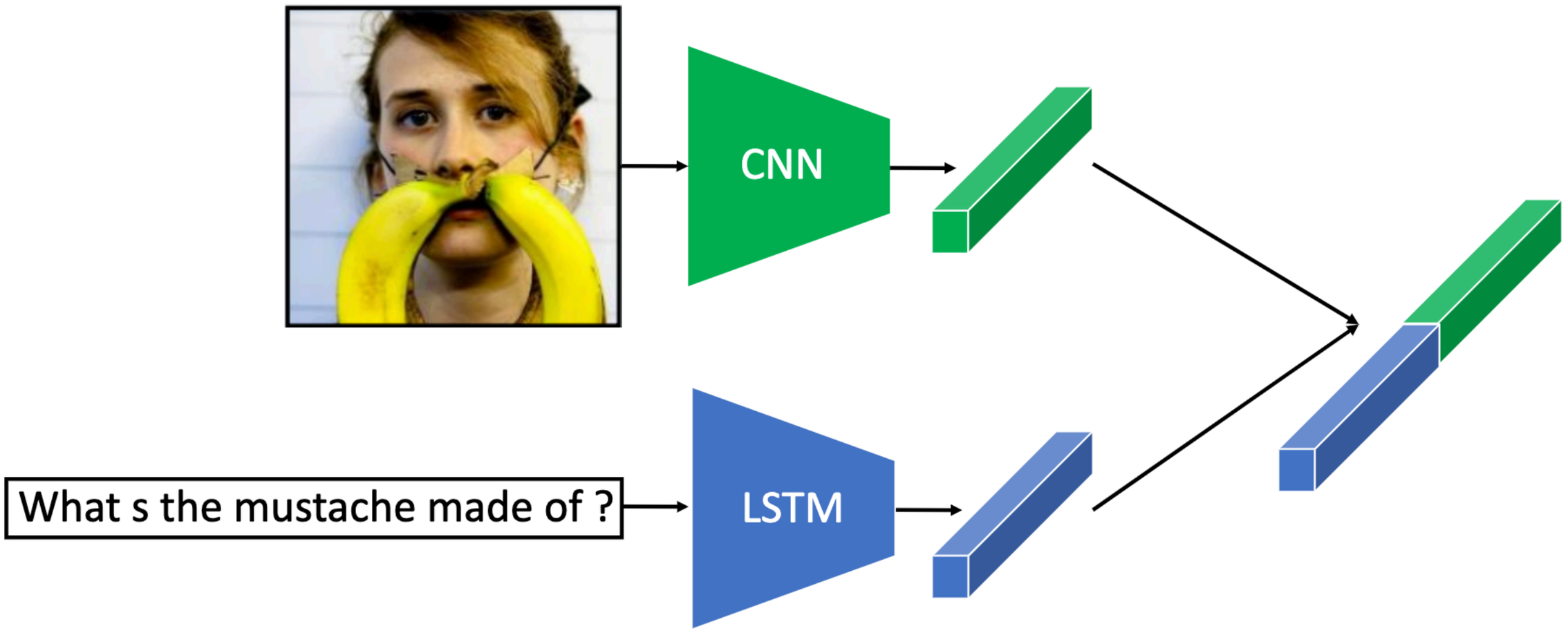


- Combine benefits of both early and late fusion mechanisms

Model-based Fusion Approaches (simple MLP approaches)

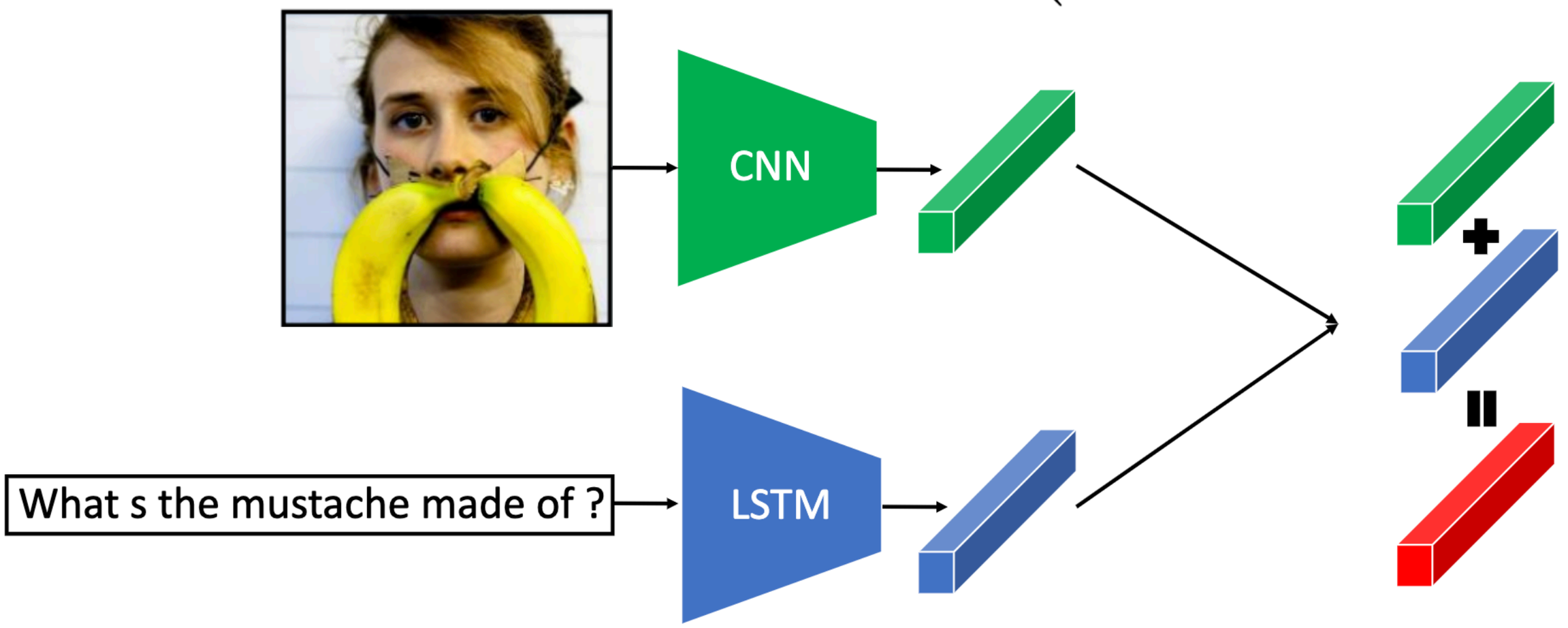
Concatenate

$$\mathbf{y} = MLP([\mathbf{x}^1, \mathbf{x}^2])$$



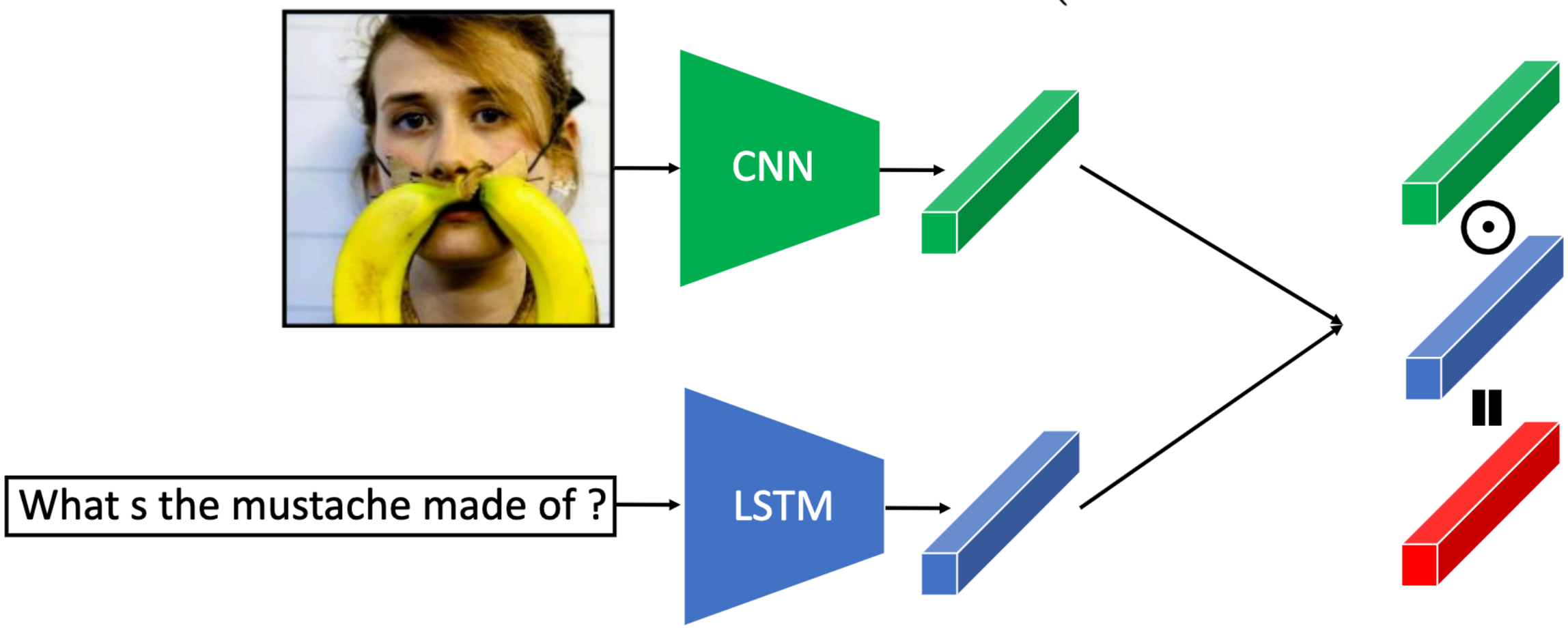
Element-wise sum

$$\mathbf{y} = W^{out}\left(W^1\mathbf{x}^1 + W^2\mathbf{x}^2\right)$$

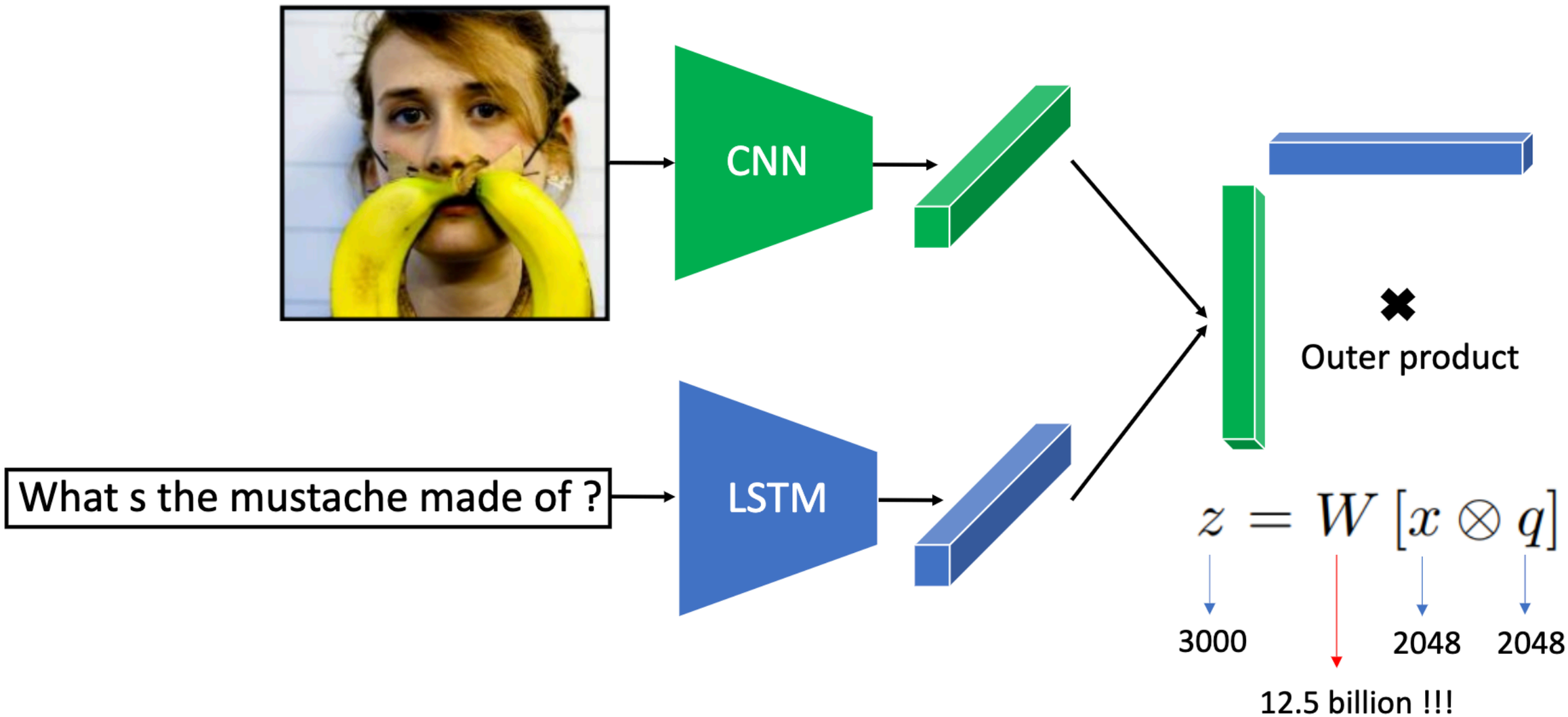


Element-wise product

$$\mathbf{y} = W^{out}\left(W^{1}\mathbf{x}^{1} \odot W^{2}\mathbf{x}^{2}\right)$$



Bilinear pooling



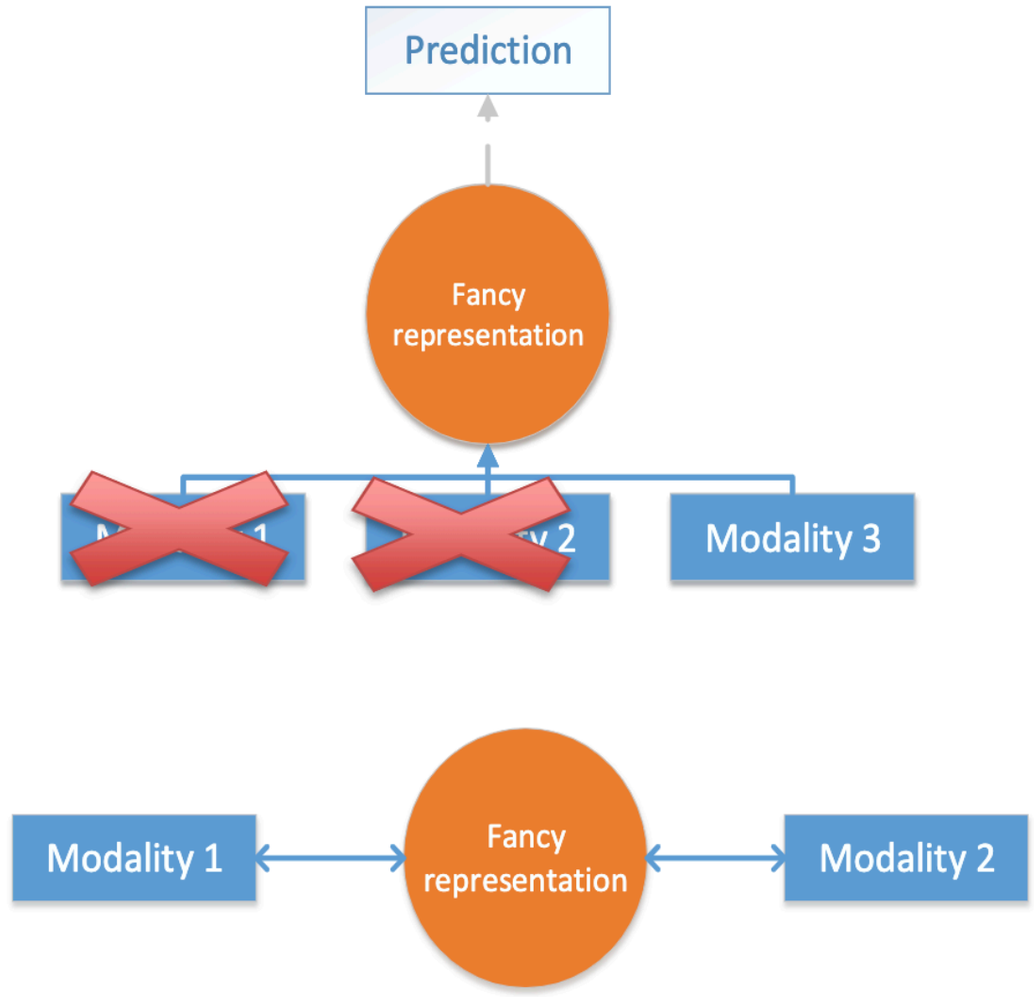
Multimodal Representation

Multimodal Represetnation

- Why do we need multimodal representations?
- Can just have unimodal ones and just fuse them
  - What if relationship is complex?
  - Doesn’t exploit joint information, especially at lower/intermediate levels

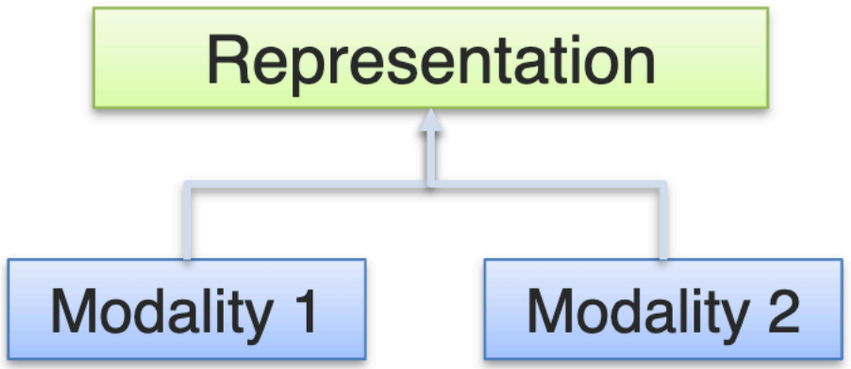
Multimodal Represetnation

- What do we want from multi-modal representation
  - Similarity in that space implies similarity in corresponding concepts
  - Useful for various discriminative tasks – retrieval, mapping, fusion etc.
  - Possible to obtain in absence of one or more modalities
  - Fill in missing modalities given others (map between modalities)



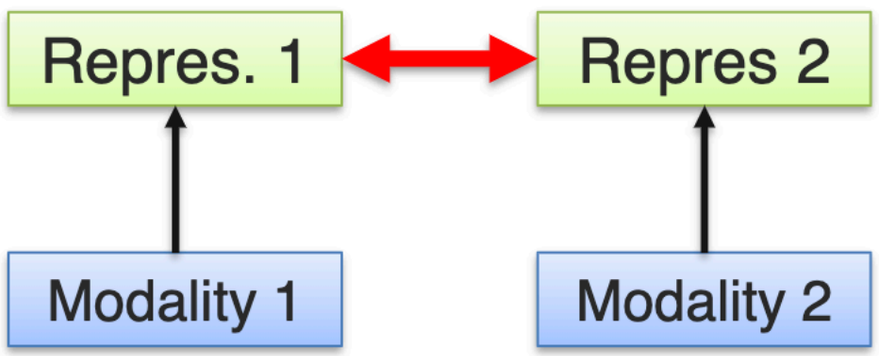
Multimodal representation types

Ⓐ Joint represenations:



- Simplest version: modality concatenation (early fusion)
- Can be learned supervised or unsupervised
- Multimodal factor analysis

Ⓑ Coordinated representations

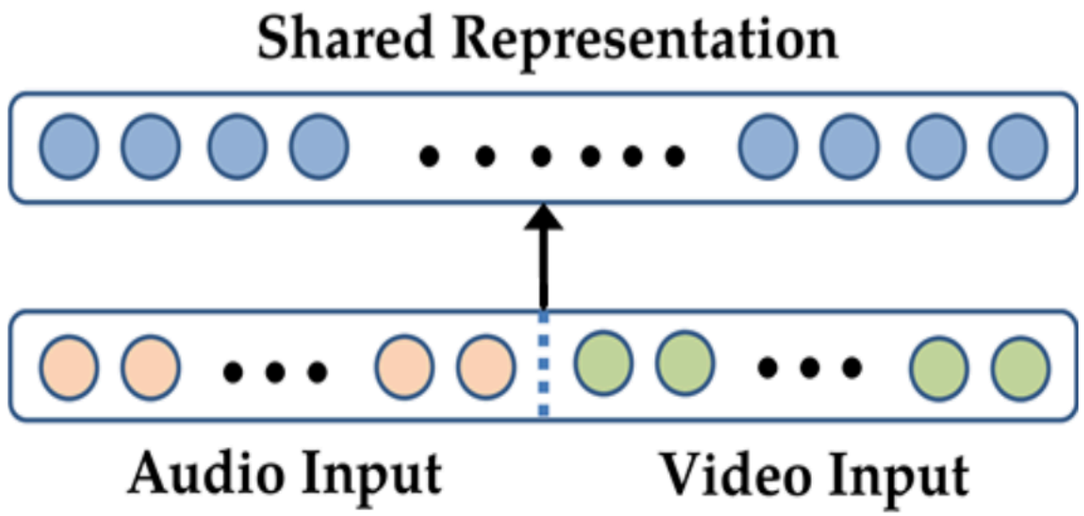


- Similarity-based methods (e.g., cosine distance)
- Structure constraints (e.g., orthogonality, sparseness)

Joint Representation

Shallow multimodal representations

- Want deep multimodal representations
  - Shallow representations do not capture complex relationships
  - Often shared layer only maps to the shared section directly



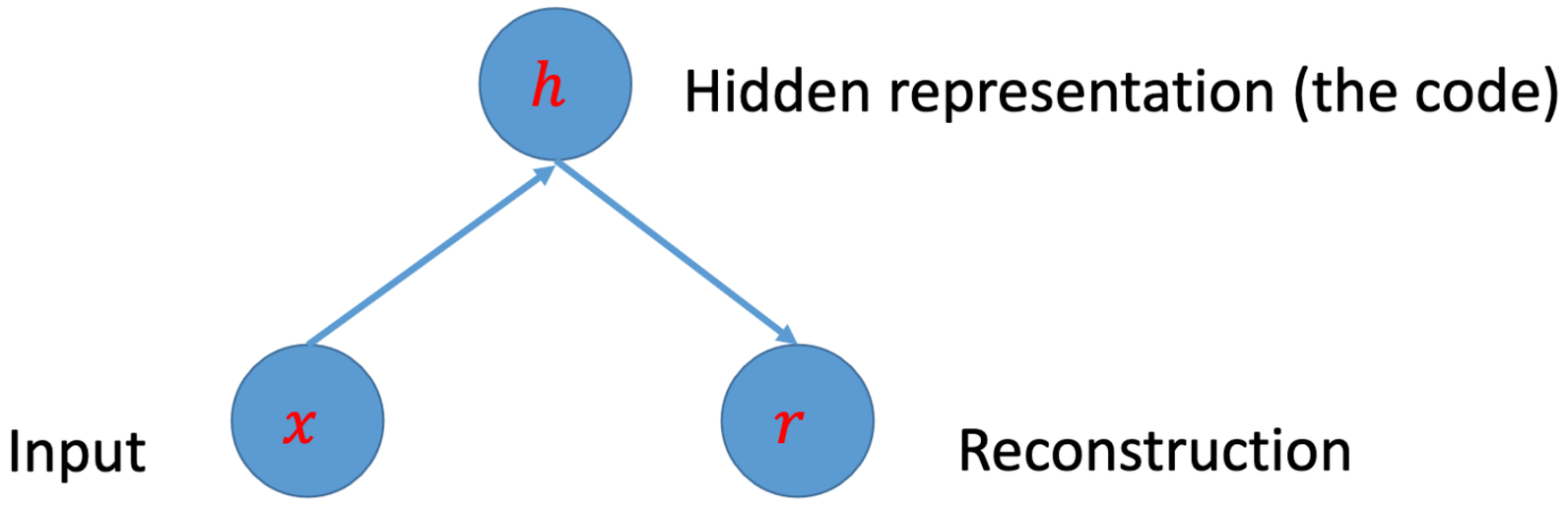
Shallow Autoencoder

Autoencoders

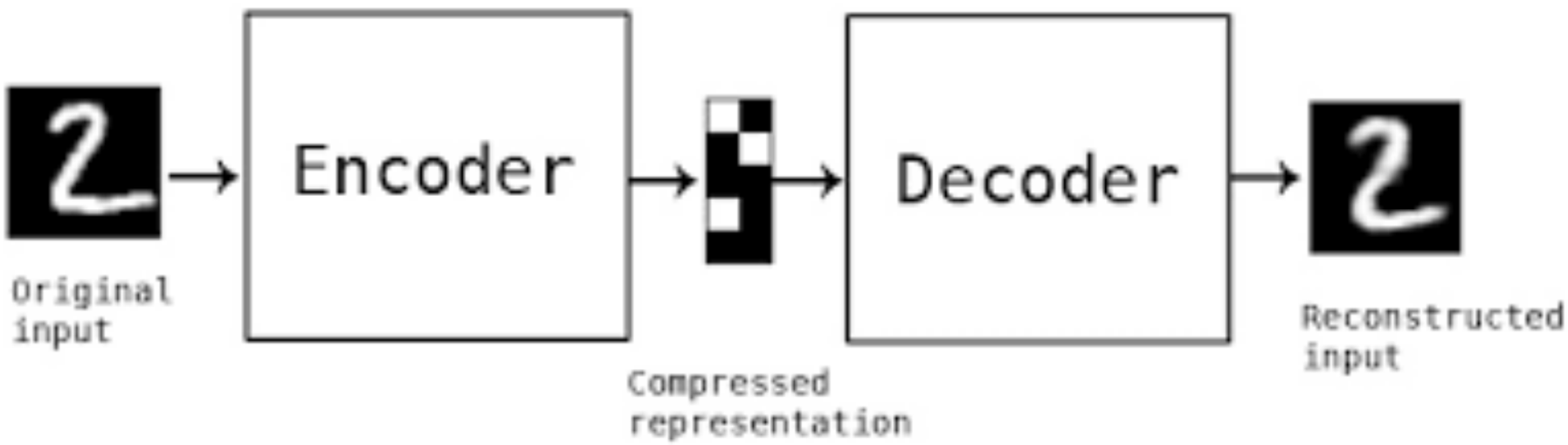
Autoencoder

- Neural networks trained to attempt to copy its input to its output

- Contain two parts:
  - Encoder: map the input to a hidden representation
  - Decoder: map the hidden representation to the output



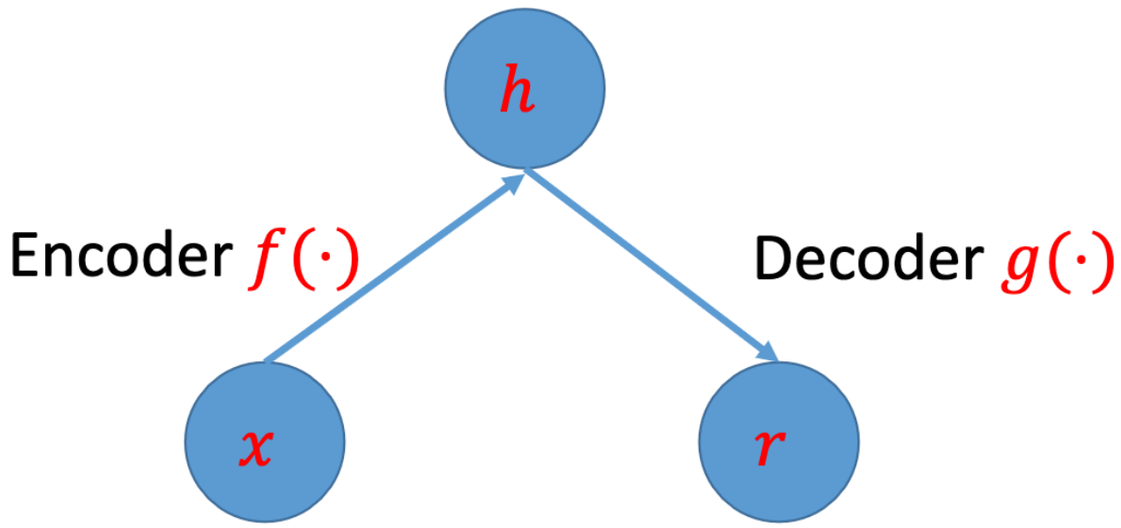
Encoder


Encoder
Decoder
Original input
Compressed representation
Reconstructed input

Autoencoder

- Autoencoder forced to select which aspects to preserve and thus hopefully can learn useful properties of the data

- Historical note: goes back to (LeCun, 1987; Bourlard and Kamp, 1988; Hinton and Zemel, 1994).

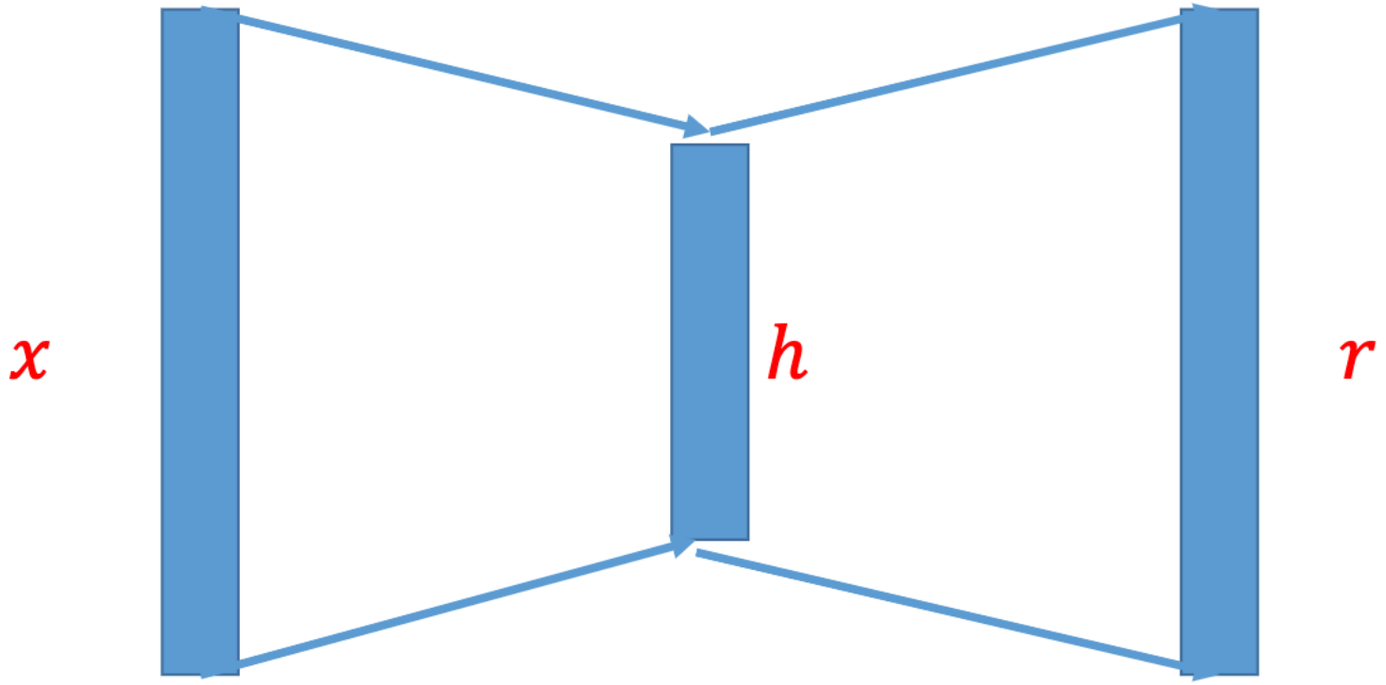


$$h = f(x), r = g(h) = g(f(x))$$

Autoencoder

- Constrain the code to have smaller dimension than the input
- Training: minimize a loss function

$$L(x, r) = L(x, g(f(x)))$$



Autoencoder

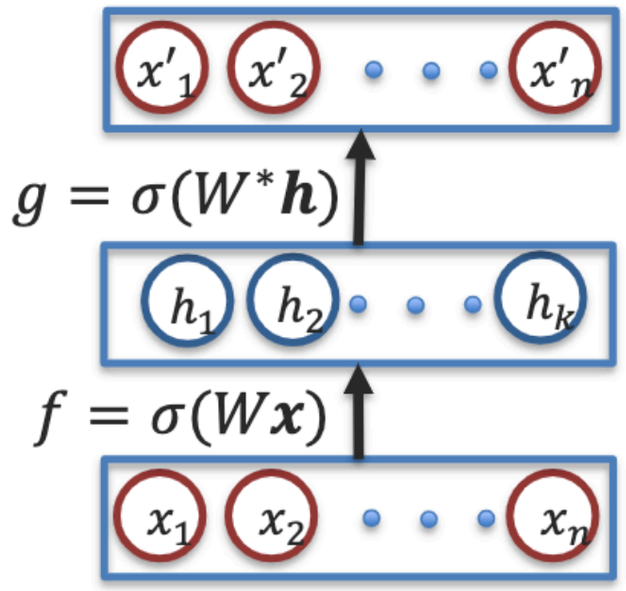
- Constrain the code to have smaller dimension than the input
- Training: minimize a loss function

$$L(x, r) = L(x, g(f(x)))$$

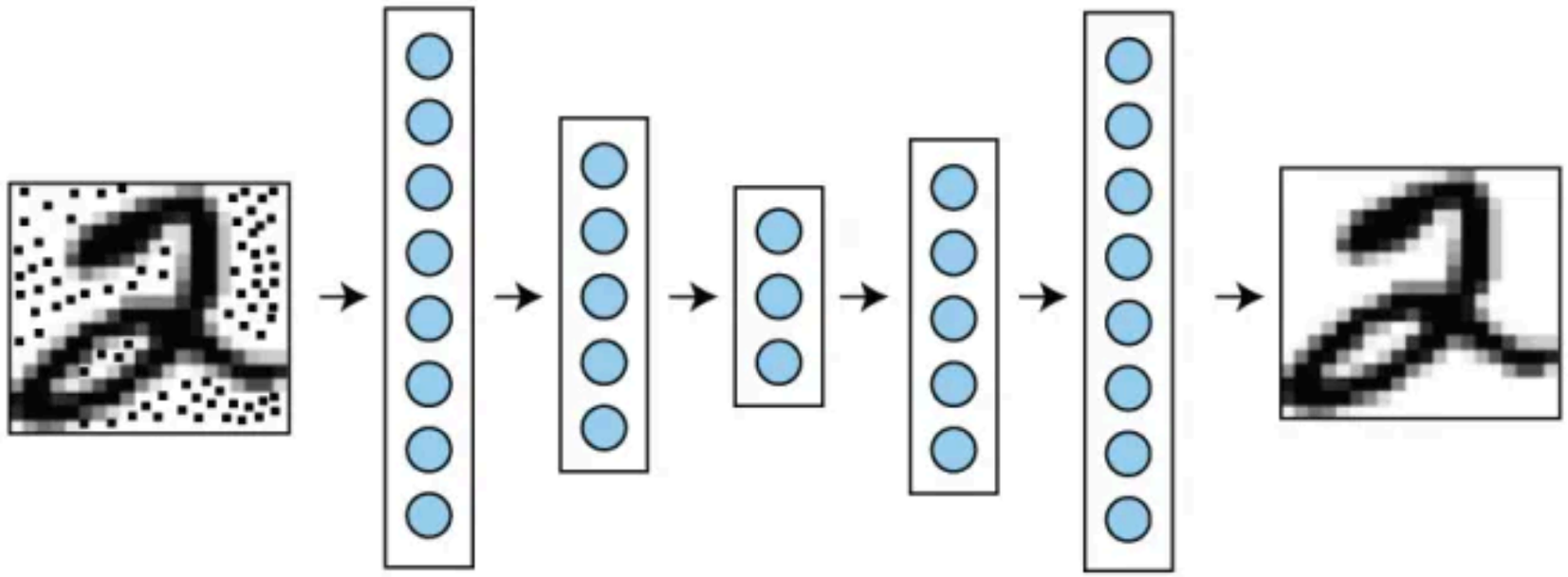
- Special case: f, g linear, L mean square error
- Reduces to Principal Component Analysis

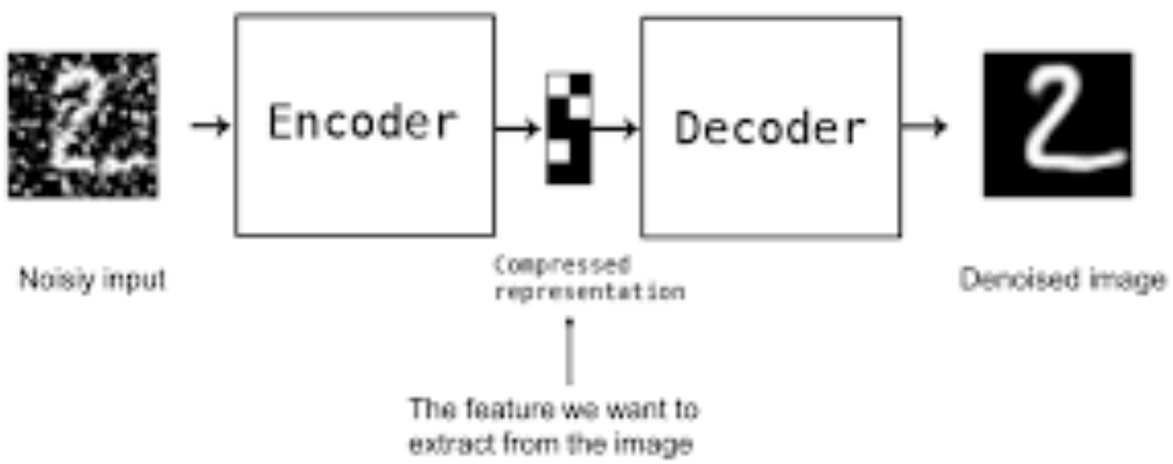
Autoencoders

- Mostly follows Neural Network structure
  - A matrix multiplication followed by a sigmoid
- Activation will depend on type of $\boldsymbol{x}$
  - Sigmoid for binary
  - Linear for real valued
- Often we use tied weights to force the sharing of weights in encoder/decoder
  - $W^* = W^T$
- word2vec is actually a bit similar to autoencoder



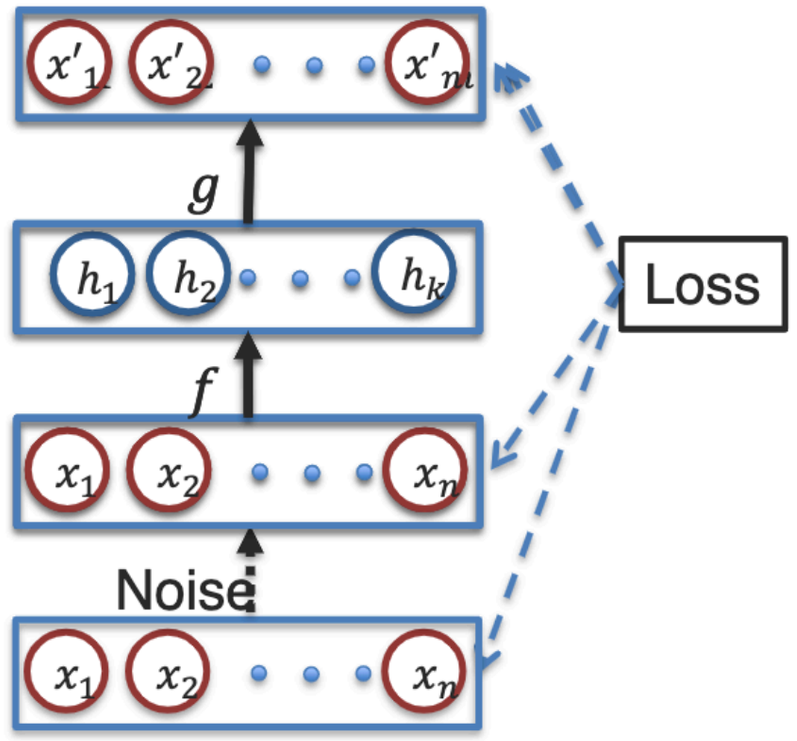
Denoising Autoencoder





Denoising Autoencoder

- Simple idea
  - Add noise to input $\boldsymbol{x}$ but learn to reconstruct original
- Leads to a more robust representation and prevents copying
- Learns what the relationship is to represent a certain $\boldsymbol{x}$
- Different noise added during each epoch



Denoising Autoencoder

- Traditional autoencoder: encourage to learn $g\big(f(\cdot)\big)$ to be identity

- Denoising : minimize a loss function

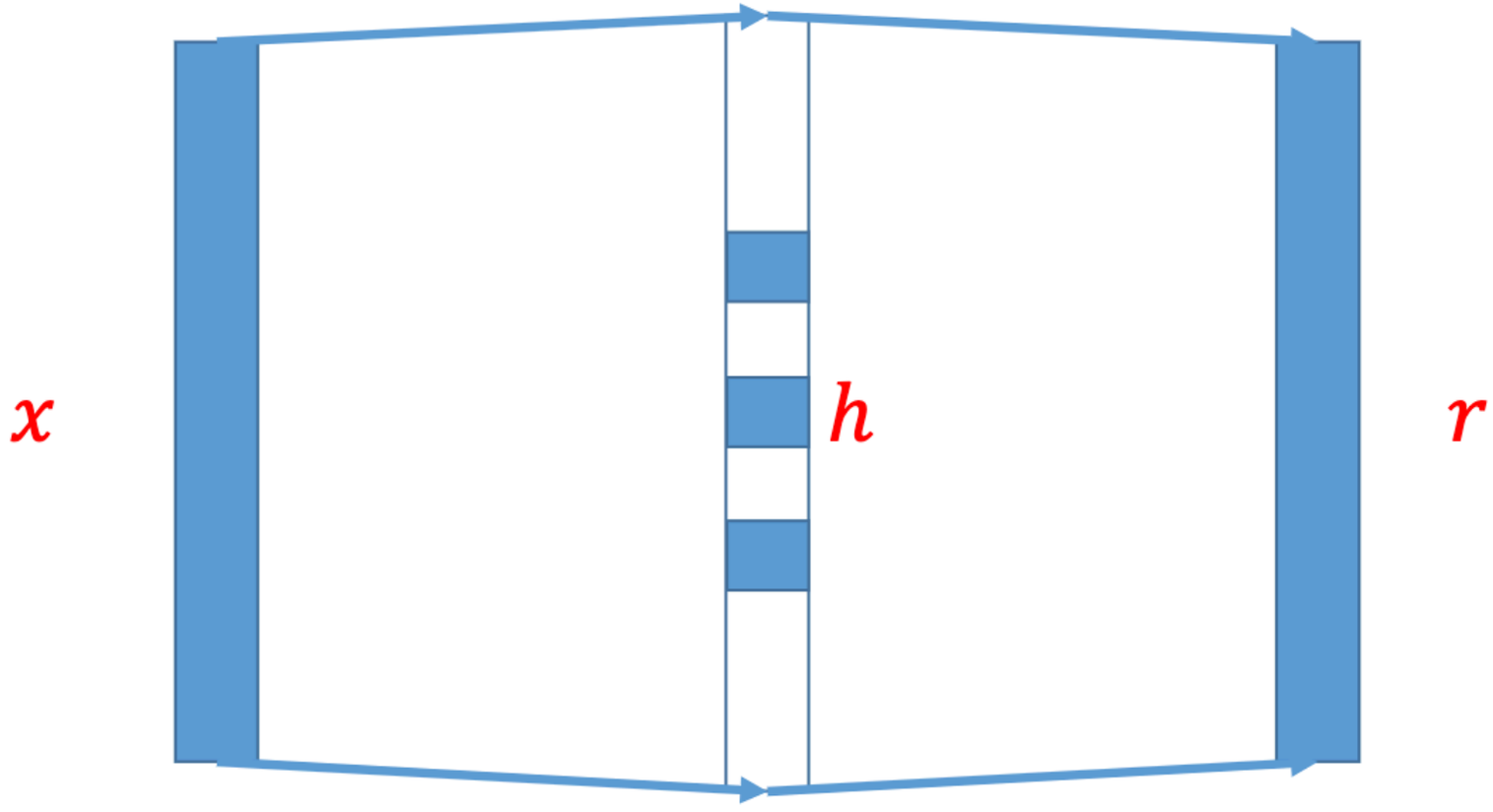
$$L(x, r) = L(x, g\big(f(\tilde{x})\big))$$

where $\tilde{x}$ is $x + noise$

Sparse autoencoder

- Constrain the code to have sparsity
- Training: minimize a loss function

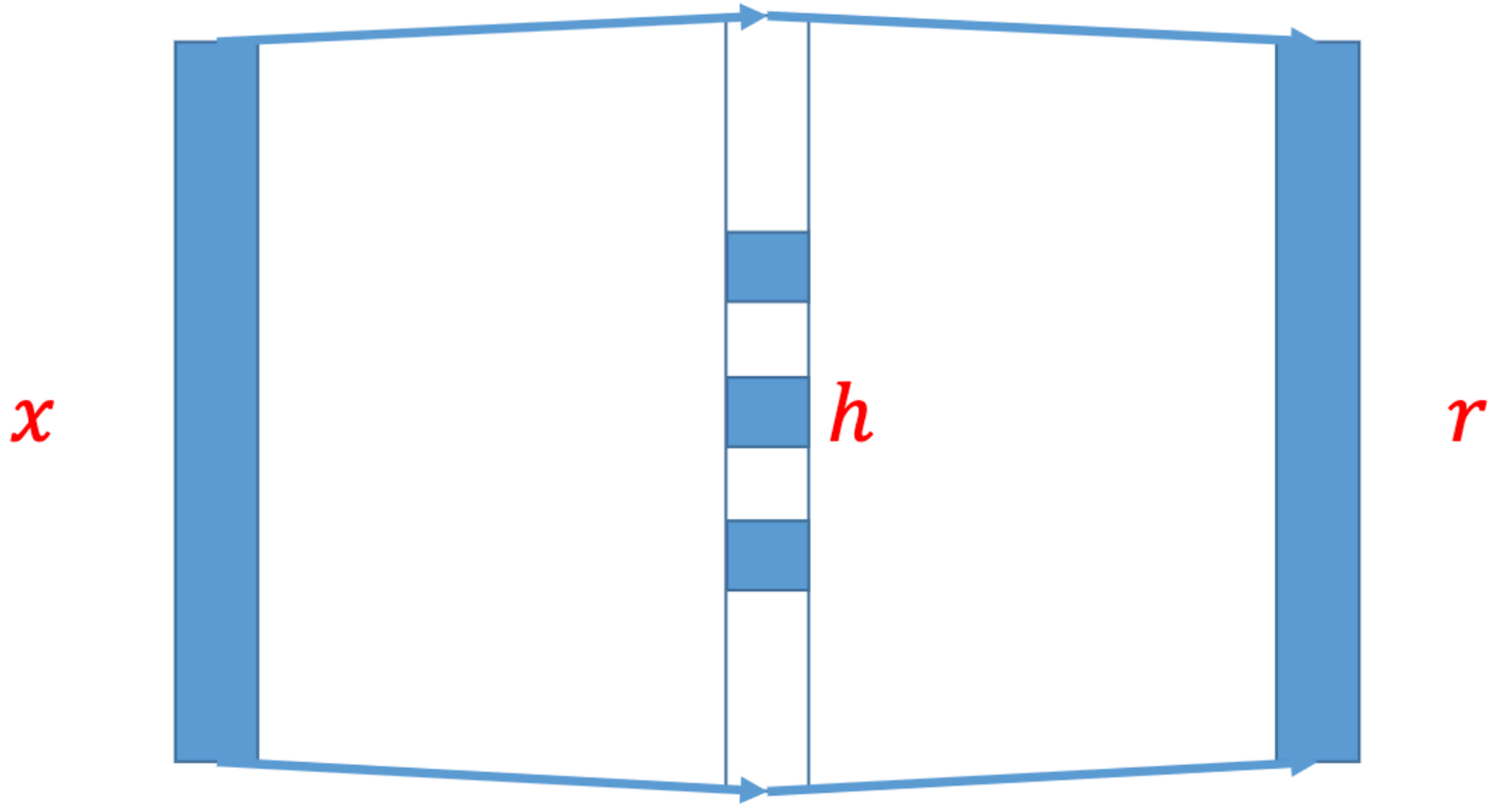
$$L_R = L(x, g(f(x))) + R(h)$$



Sparse autoencoder

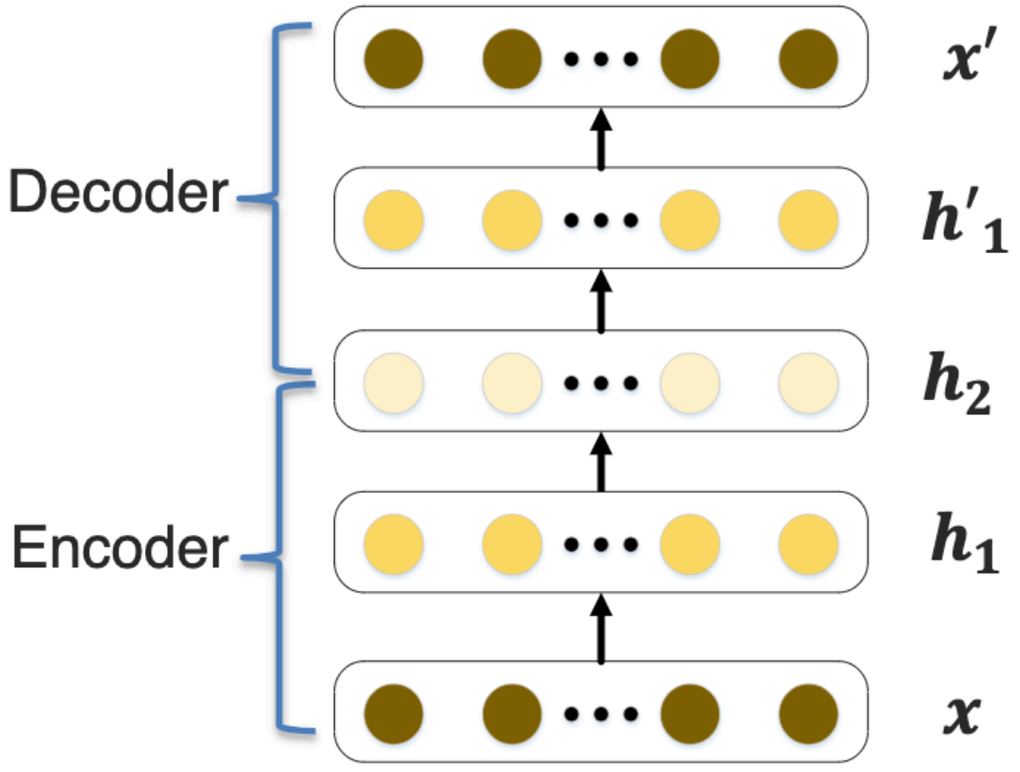
- Constrain the code to have sparsity
- Training: minimize a loss function

$$L_R = L(x, g(f(x))) + \lambda |h|_1$$



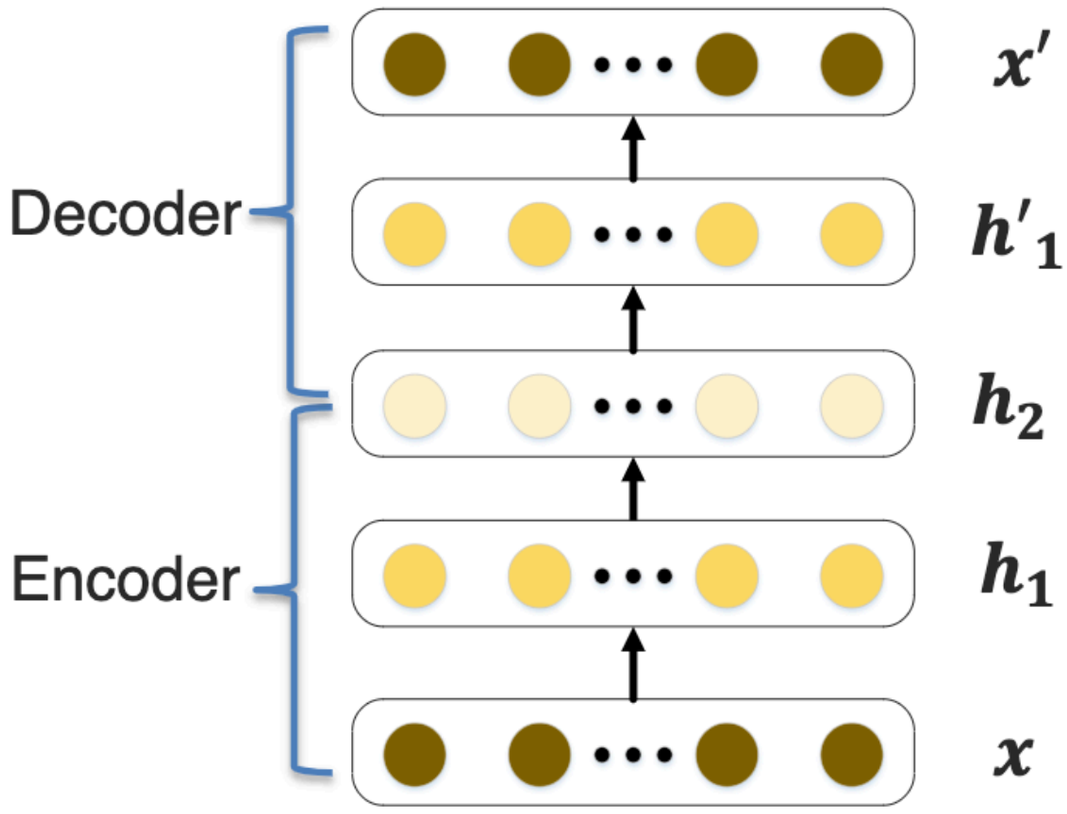
Stacked Autoencoders

- Reconstruct using previously learned decoders mappings
- Fine-tune the full network end-to-end



Stacked denoising autoencoders

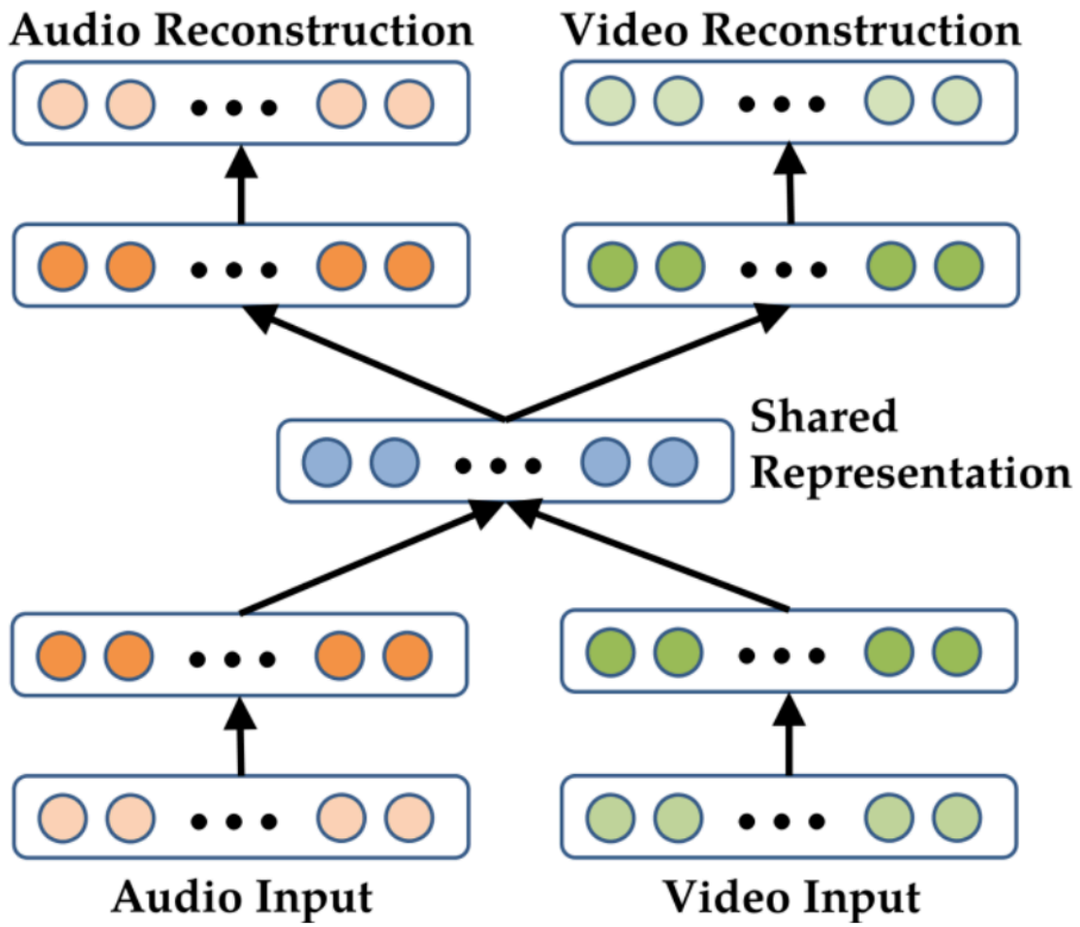
- Can extend this to a denoising model
- Add noise when training each of the layers
  - Often with increasing amount of noise per layer
  - 0.1 for first, 0.2 for second, 0.3 for third



Multimodal Autoencoders

Deep Multimodal autoencoders

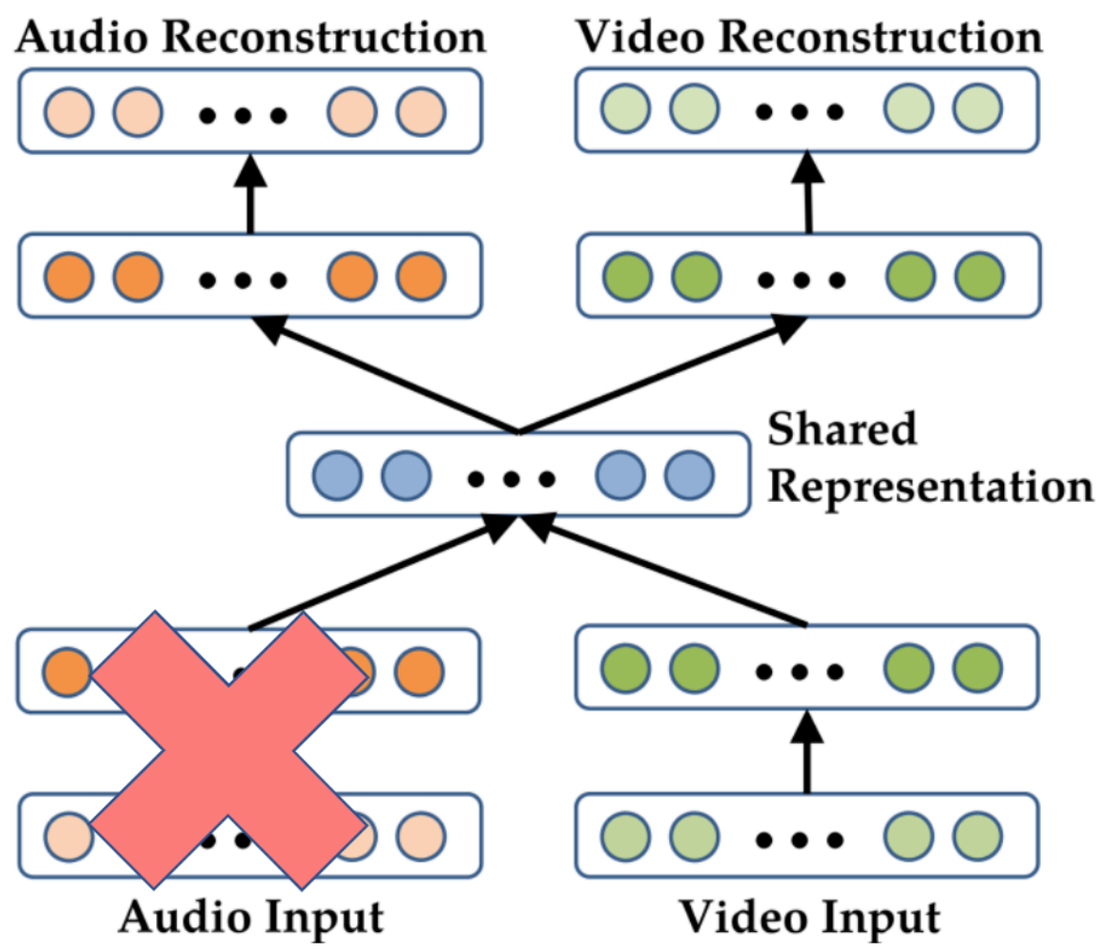
- A deep representation learning approach
- A bimodal auto-encoder
  - Used for Audio-visual speech recognition



[Ngiam et al., Multimodal Deep Learning, 2011]

Deep Multimodal autoencoders training

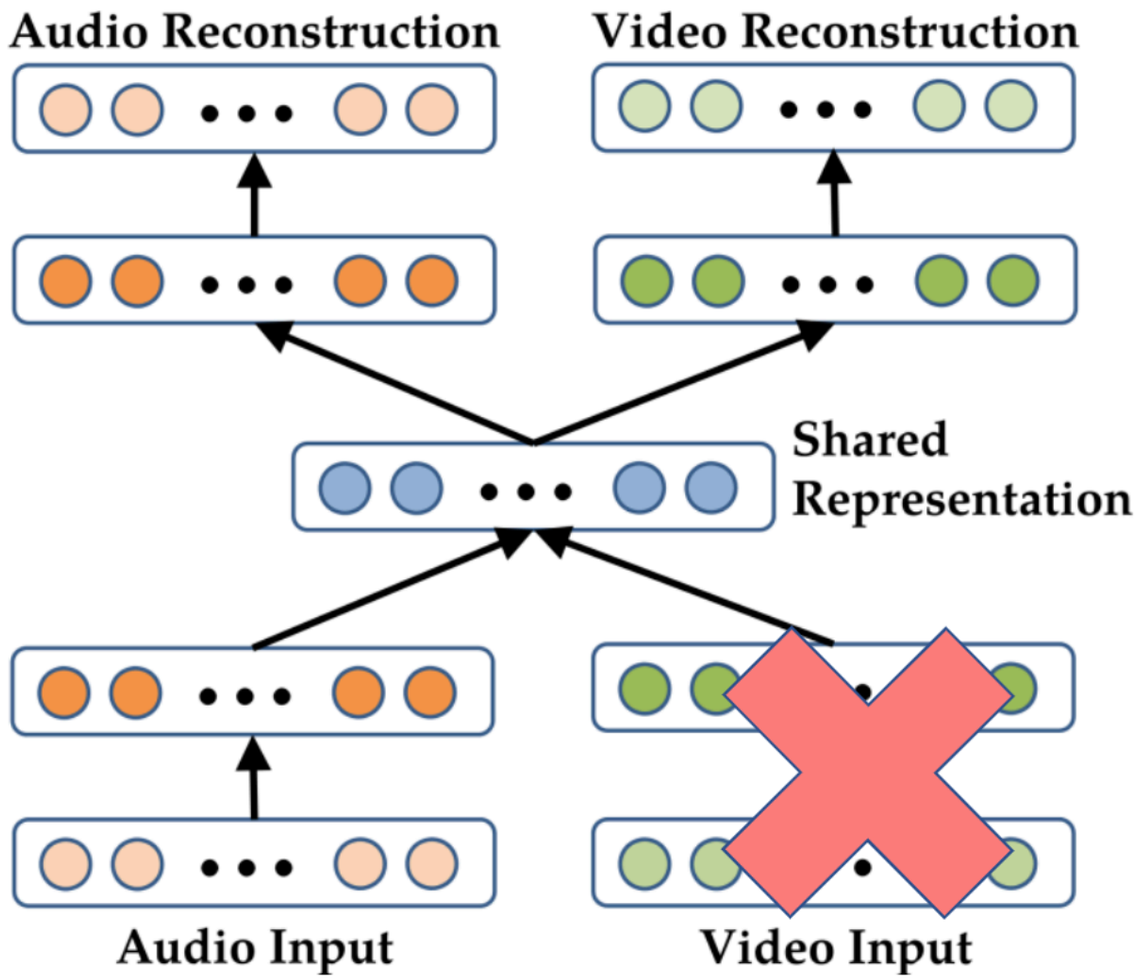
- Individual modalities can be pre-trained
  - Denoising Autoencoders
- To train the model to reconstruct the other modality
  - Use both
  - Remove audio



Deep Multimodal autoencoders training

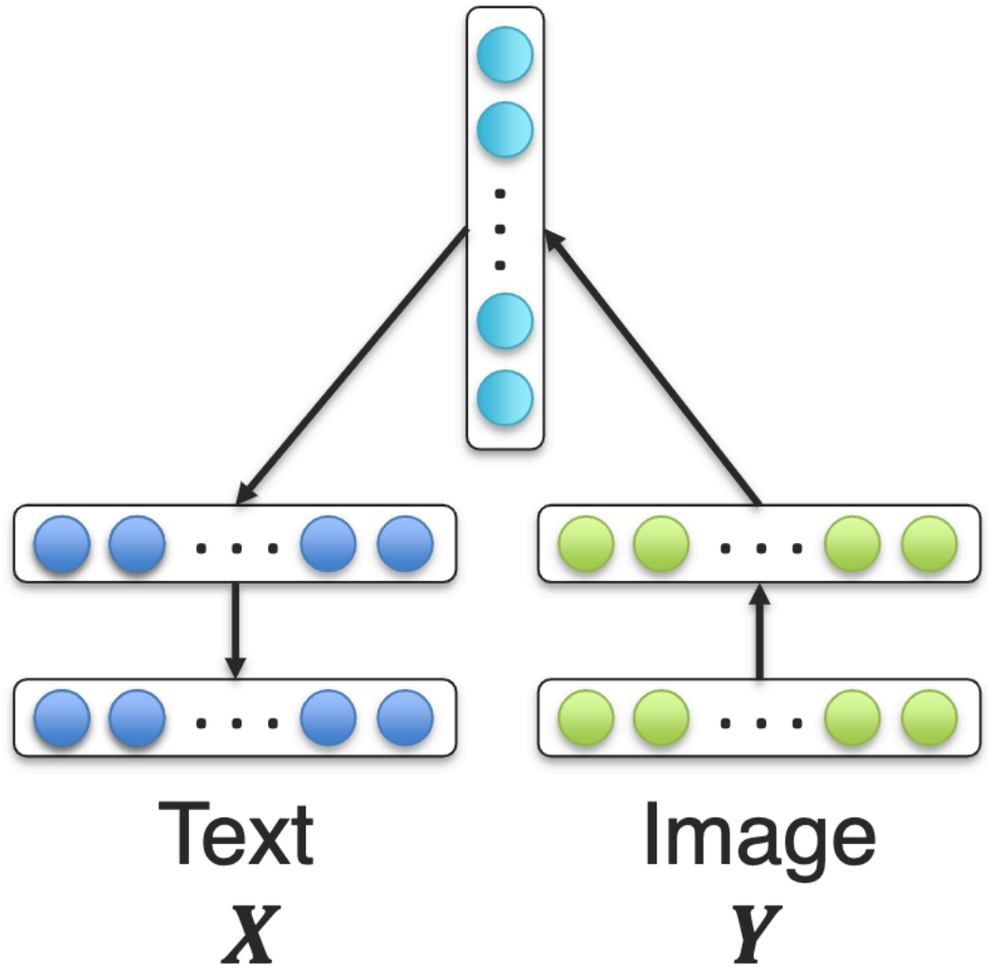
- Individual modalities can be pre-trained
  - RBMs
  - Denoising Autoencoders

- To train the model to reconstruct the other modality
  - Use both
  - Remove audio
  - Remove video



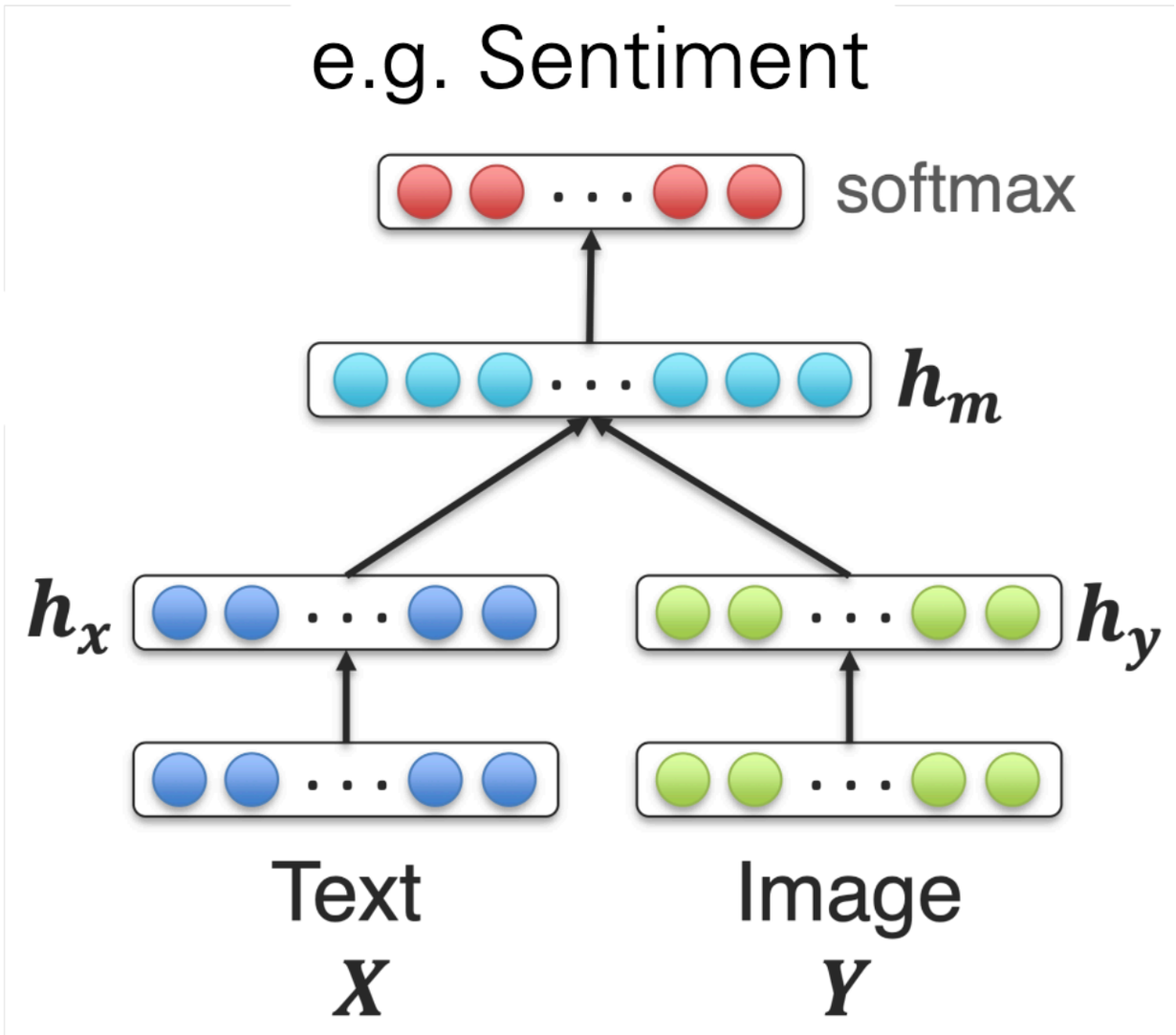
Multimodal Encoder-Decoder

- Visual modality often encoded using CNN
- Language modality will be decoded using LSTM
  - A simple multilayer perceptron will be used to translate from visual (CNN) to language (LSTM)

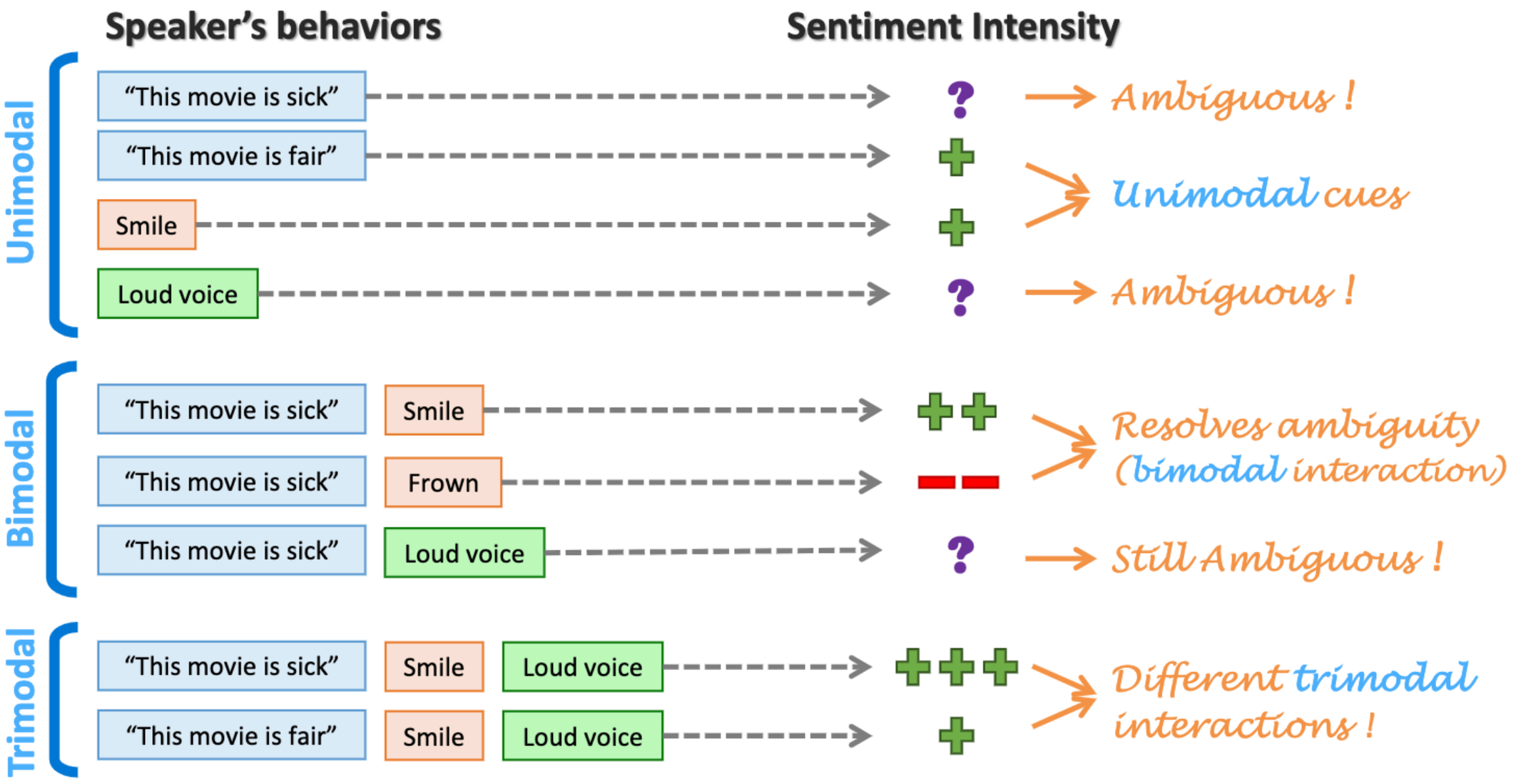


Multimodal Joint Representation

- For supervised learning tasks
- Joining the unimodal representations:
  - Simple concatenation
  - Element-wise multiplication or summation
  - Multilayer perceptron
- How to explicitly model both unimodal and bimodal interactions?



Unimodal, Bimodal and Trimodal Interactions



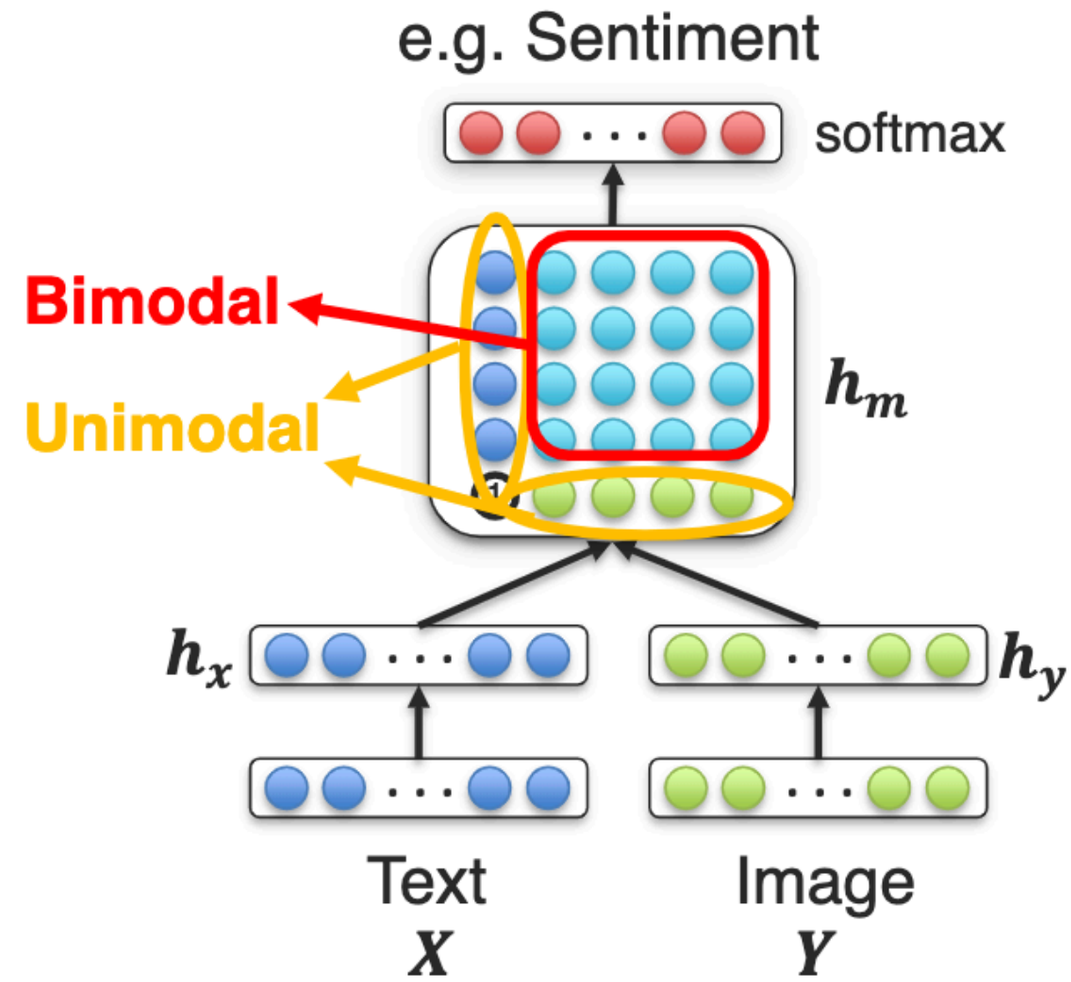
Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} = \begin{bmatrix} h_x & h_x \otimes h_y \\ 1 & h_y \end{bmatrix}$$

Important !

[Zadeh, Jones and Morency, **EMNLP 2017**]



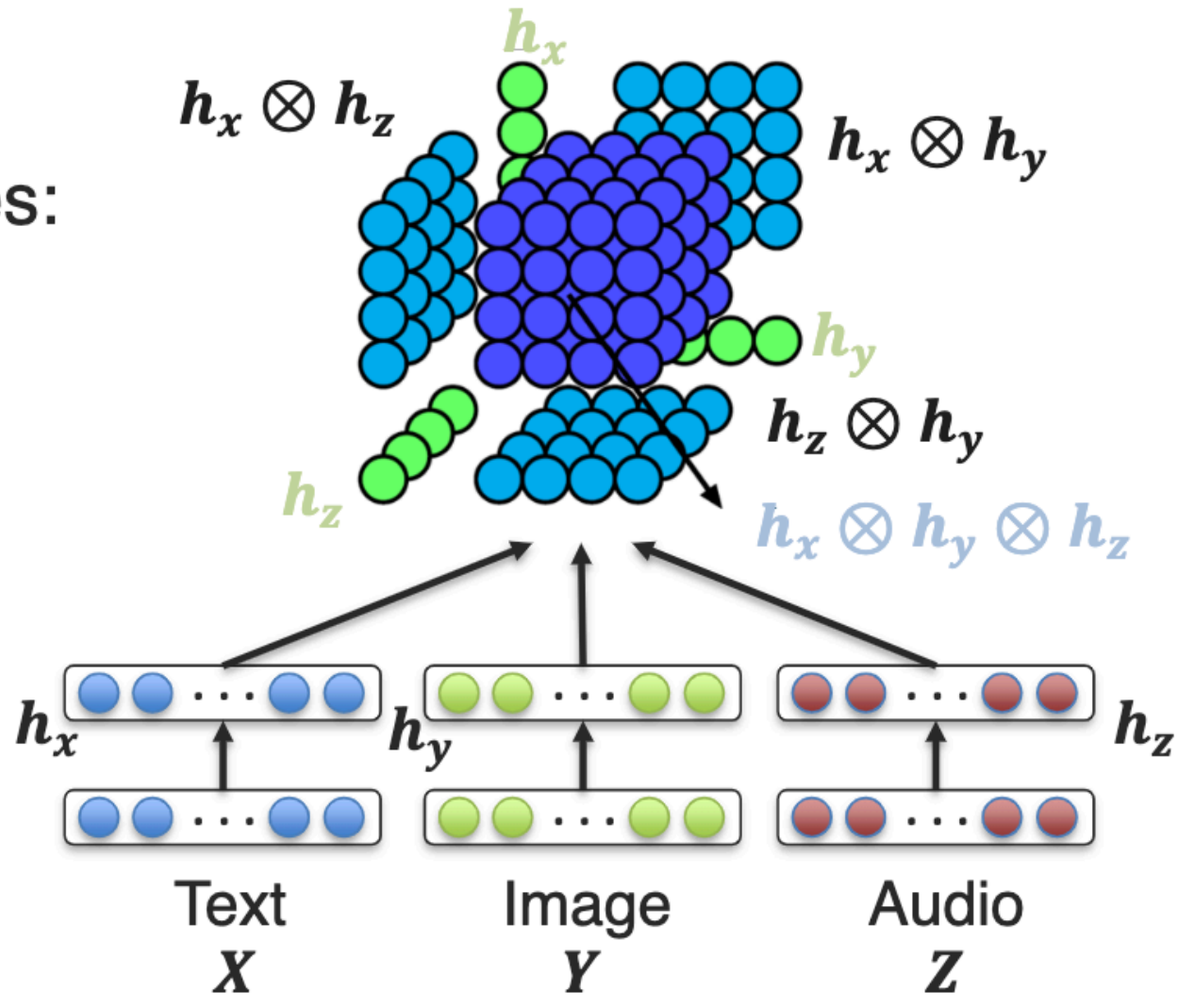
Multimodal Tensor Fusion Network (TFN)

Can be extended to three modalities:

$$h_m = \begin{bmatrix} h_x \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_y \\ 1 \end{bmatrix} \otimes \begin{bmatrix} h_z \\ 1 \end{bmatrix}$$

Explicitly models unimodal, bimodal and trimodal interactions !

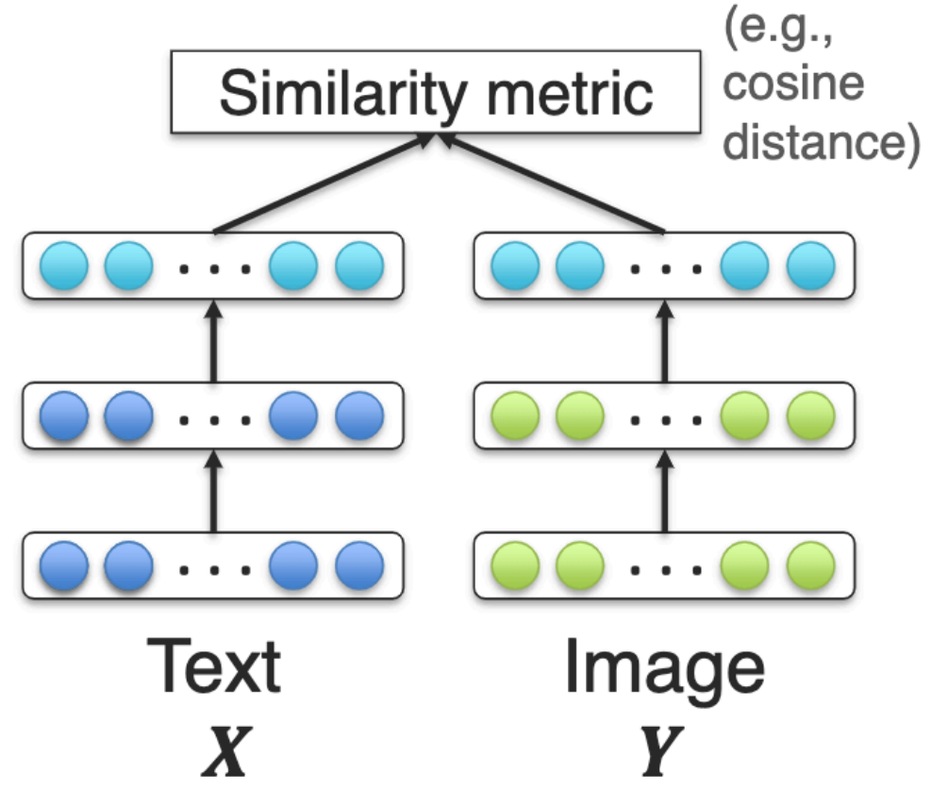
[Zadeh, Jones and Morency, **EMNLP 2017**]



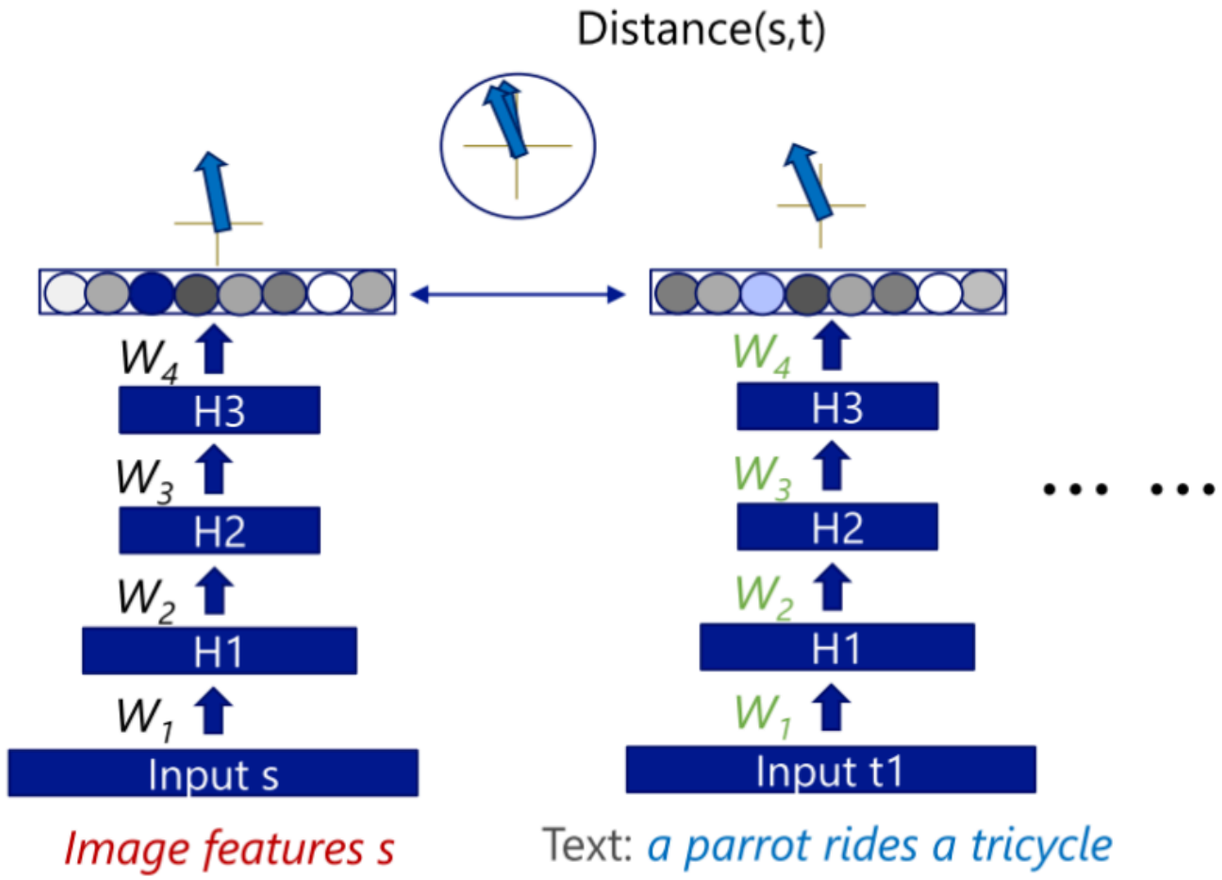
Coordinated Representations

Coordinated Multimodal Representations

Learn (unsupervised) two or more coordinated representations from multiple modalities. A loss function is defined to bring closer these multiple representations.



Coordinated Multimodal Embeddings

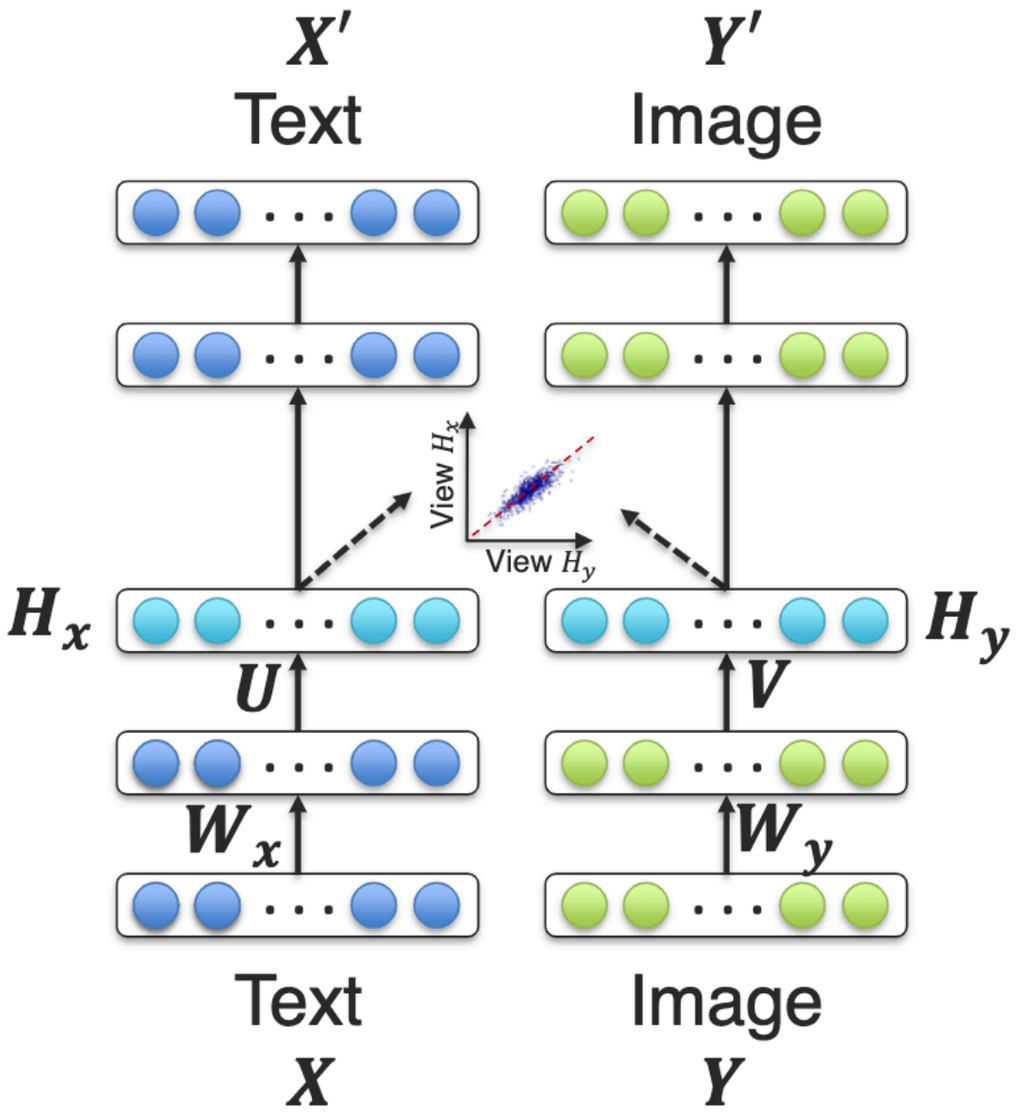


[Huang et al., Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, 2013]

Deep Canonically Correlated Autoencoders (DCCAE)

Jointly optimize for DCCA and autoencoders loss functions

- A trade-off between multi-view correlation and reconstruction error from individual views



Wang et al., ICML 2015

Recap: Multimodal representations

- Joint representations
  - Project modalities to the same space
  - Use when all the modalities are present during test time
  - Suitable for multimodal fusion

- Coordinated representations
  - Project modalities to their own coordinated space
  - Use when only one of the modalities is present during test- time
  - Suitable for multimodal translation
  - Good for retrieval

Conferences focusing on MMfusion

- **ACMMM**: ACM multimedia
https://www.acmmm.org/2020/

- **ICMI**: ACM International Conference on Multimodal Interaction
http://icmi.acm.org/2019/