Language Technologies Institute

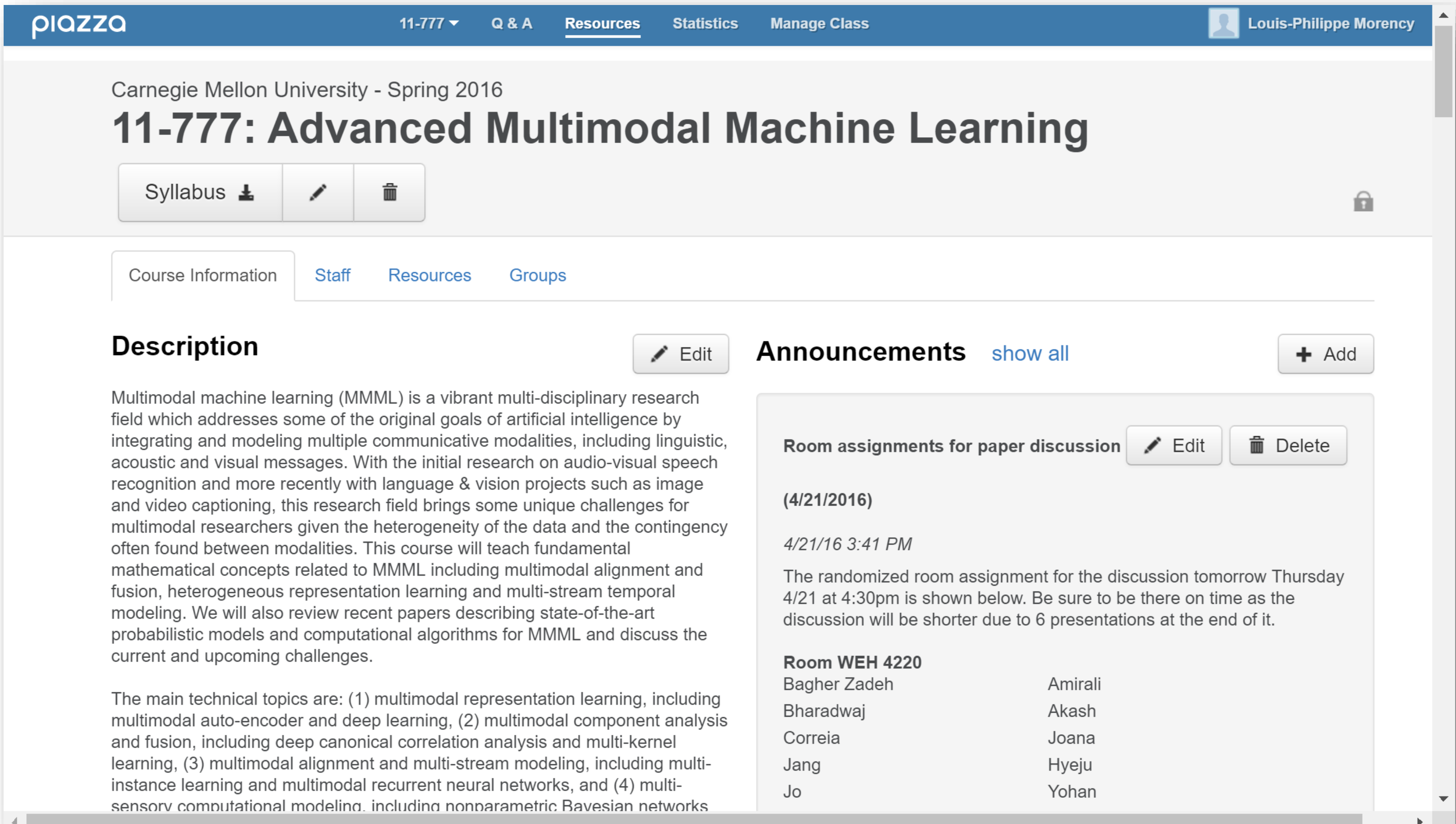
Carnegie Mellon University

Multimodal Machine Learning

Louis-Philippe (LP) Morency

CMU Multimodal Communication and Machine Learning Laboratory [MultiComp Lab]

CMU Course 11-777: Multimodal Machine Learning

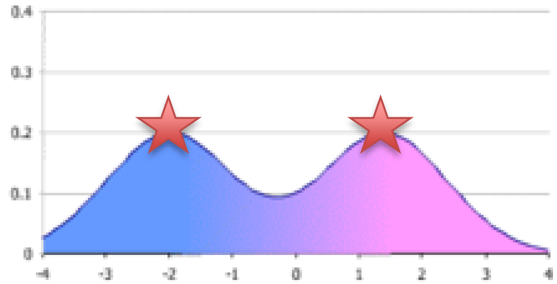


Lecture Objectives

- What is Multimodal?
- Multimodal: Core technical challenges
  - Representation learning, translation, alignment, fusion and co-learning
- Multimodal representation learning
  - Multimodal tensor representation
- Implicit Alignment
  - Temporal attention
- Fusion and temporal modeling
  - Multi-view LSTM and memory-based fusion

What is Multimodal?

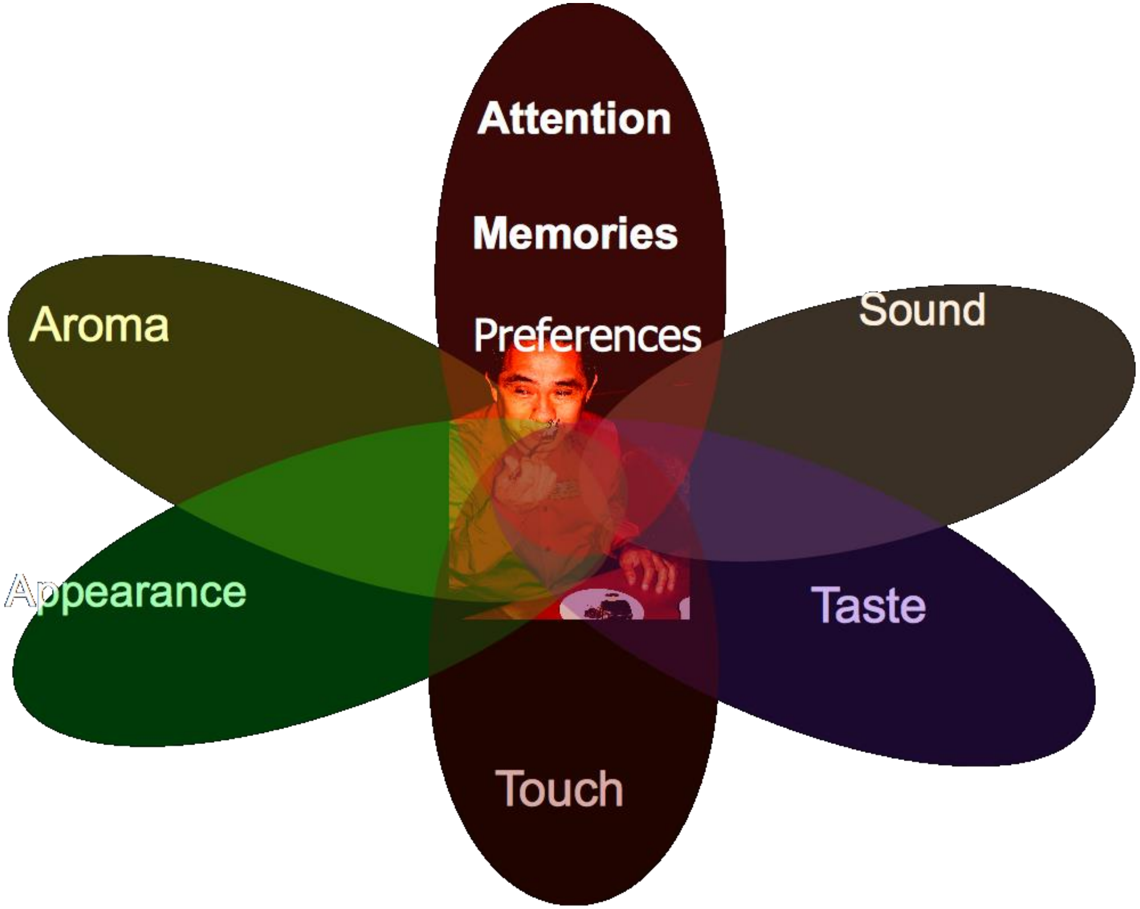
What is Multimodal?



Multimodal distribution

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

What is Multimodal?



Sensory Modalities

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

What is Multimodal?

Modality

The way in which something happens or is experienced.

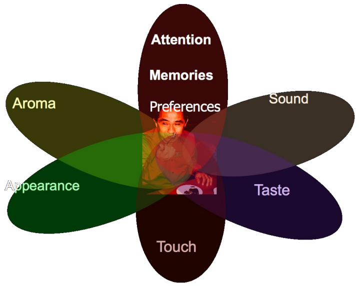
- Modality refers to a certain type of information and/or the representation format in which information is stored.
- Sensory modality: one of the primary forms of sensation, as vision or touch; channel of communication.

Medium ("middle")

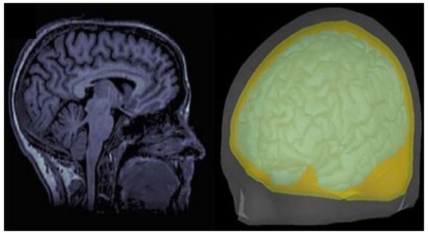
A means or instrumentality for storing or communicating information; system of communication/transmission.

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

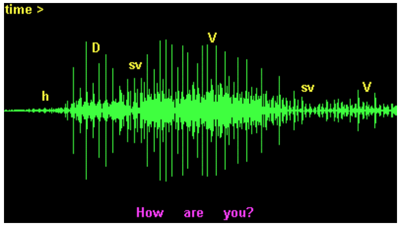
Multiple Communities and Modalities



Psychology



Medical

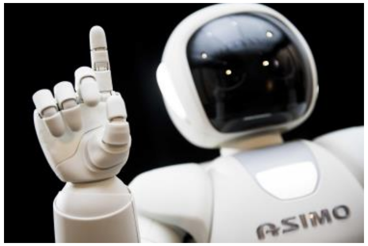


Speech

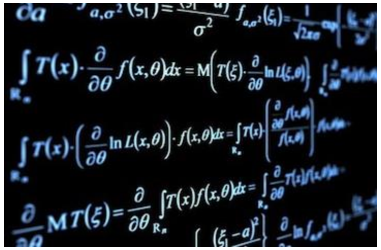
Vision

Language

Multimedia



Robotics



Learning

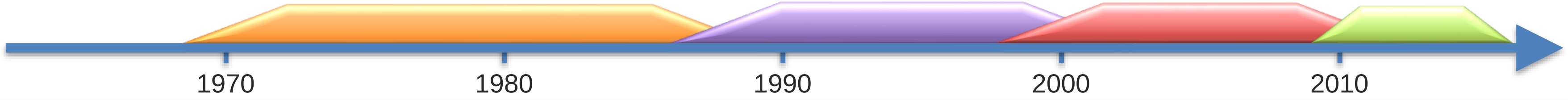
Examples of Modalities

- 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"

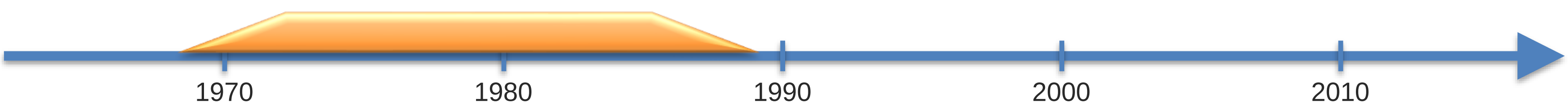
Four eras of multimodal research

- 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 …)
  - Main focus of this tutorial



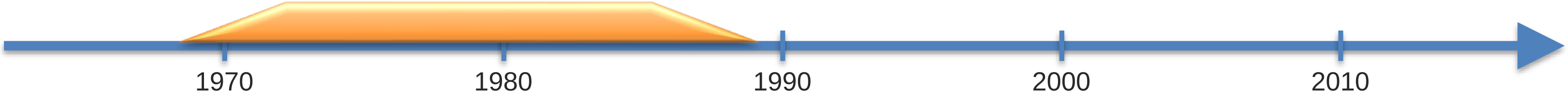
The McGurk Effect (1976)

[Hearing lips and seeing voices – Nature]



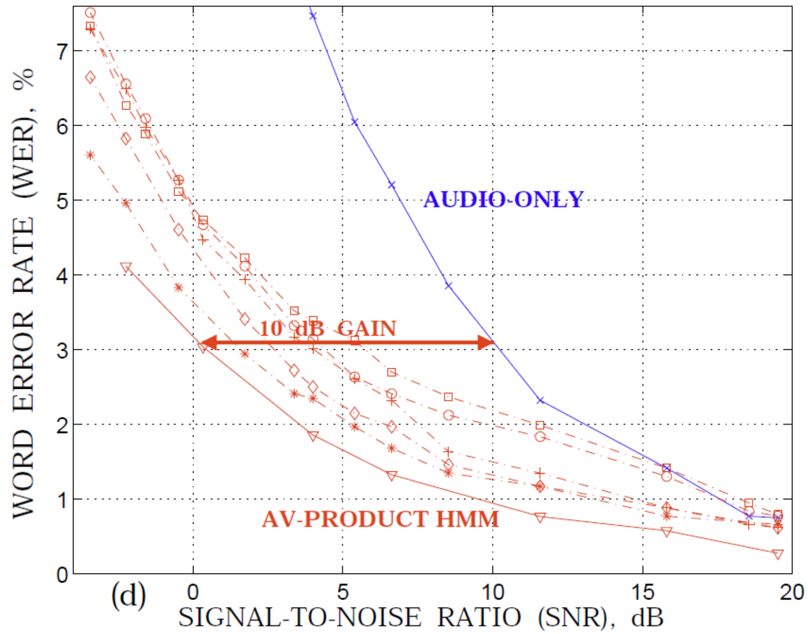
The McGurk Effect (1976)

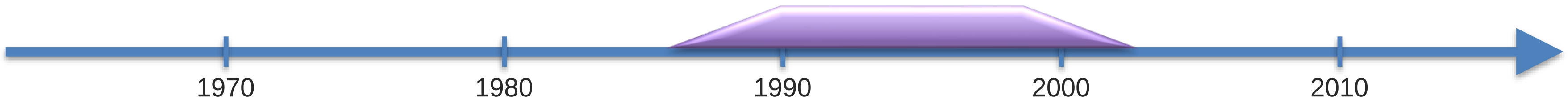
[Hearing lips and seeing voices – Nature]



➢ The “Computational” Era(Late 1980s until 2000)

1) Audio-Visual Speech Recognition (AVSR)





Core Technical Challenges

Core Challenges in “Deep” Multimodal ML

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

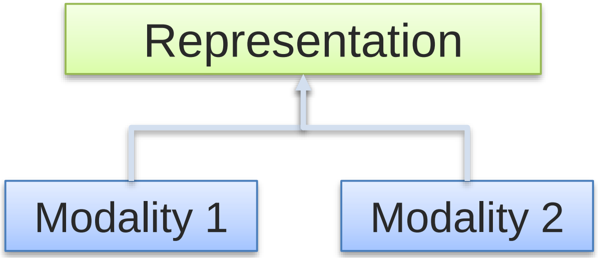
- ☑ **5 core challenges**
- ☑ **37 taxonomic classes**
- ☑ **253 referenced citations**

These challenges are non-exclusive.

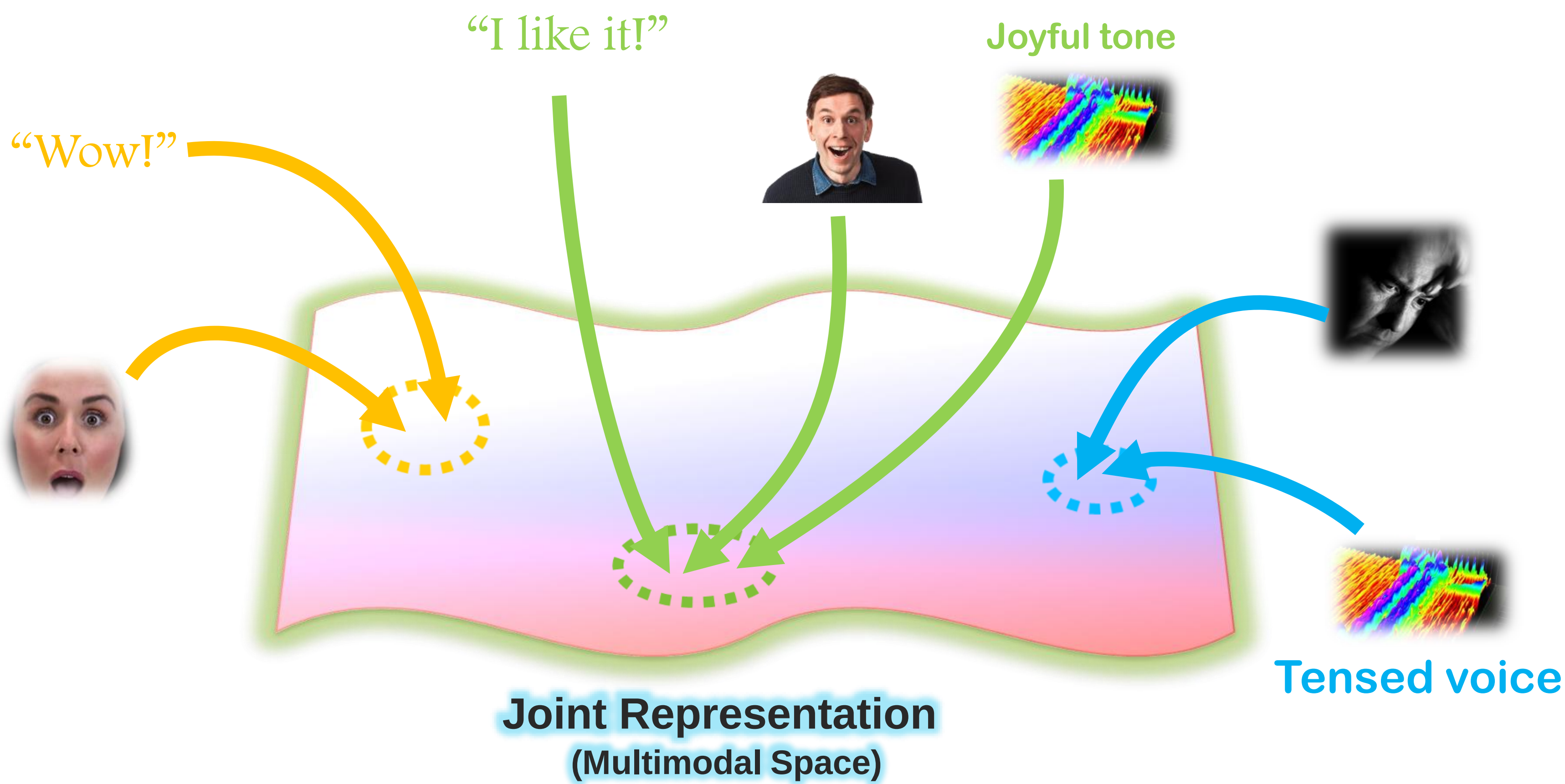
Core Challenge 1: Representation

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

(A) Joint representations:



Joint Multimodal Representation



Joint Multimodal Representations

Audio-visual speech recognition

[Ngiam et al., ICML 2011]

- Bimodal Deep Belief Network

Image captioning

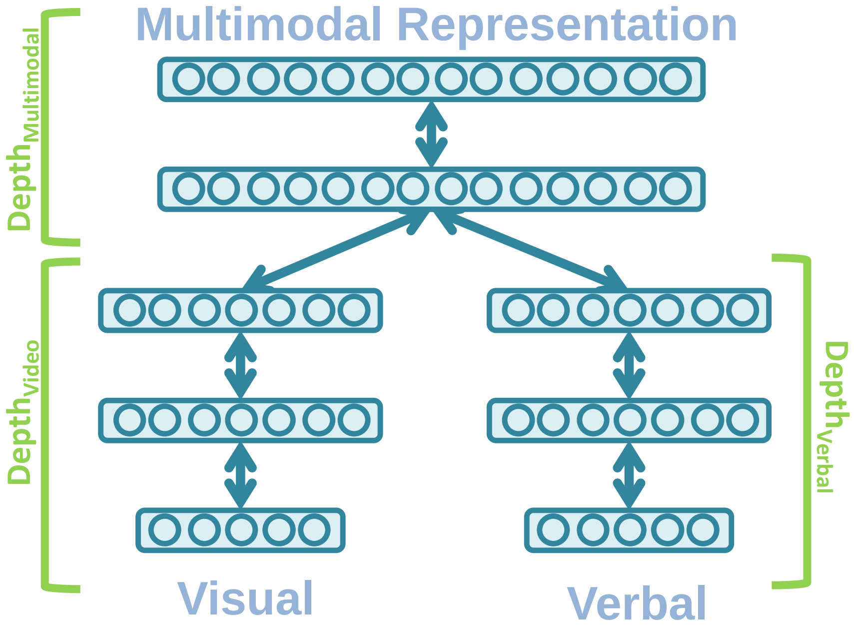
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

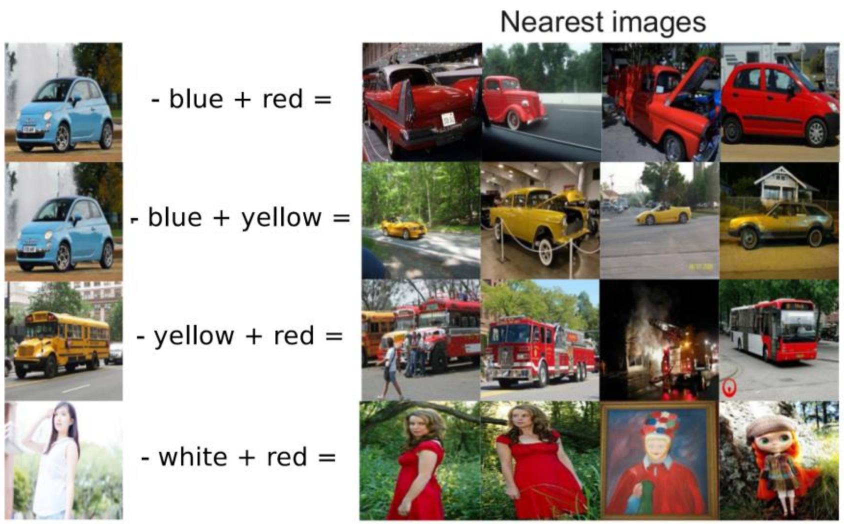
Audio-visual emotion recognition

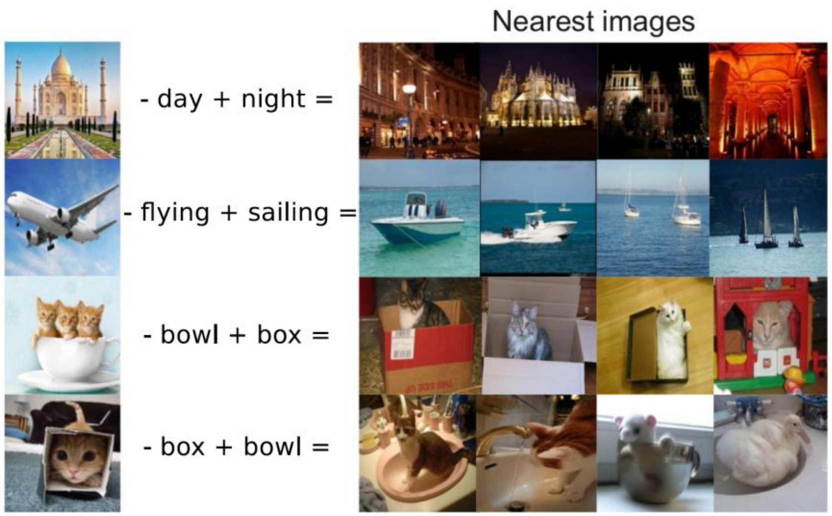
[Kim et al., ICASSP 2013]

- Deep Boltzmann Machine



Multimodal Vector Space Arithmetic



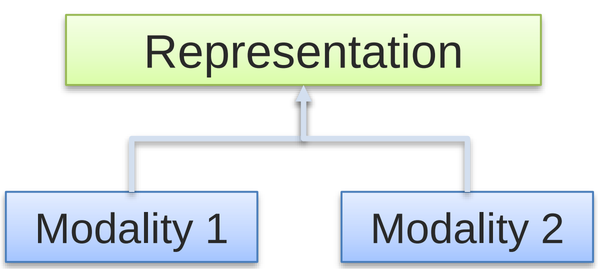


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

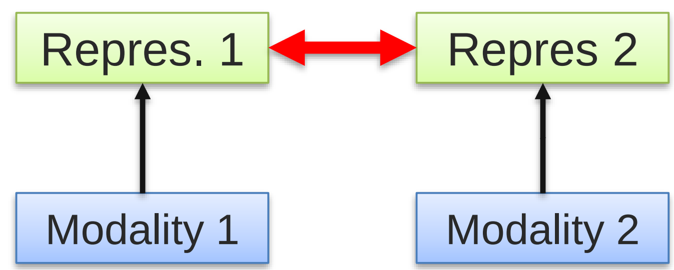
Core Challenge 1: Representation

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

Ⓐ Joint representations:



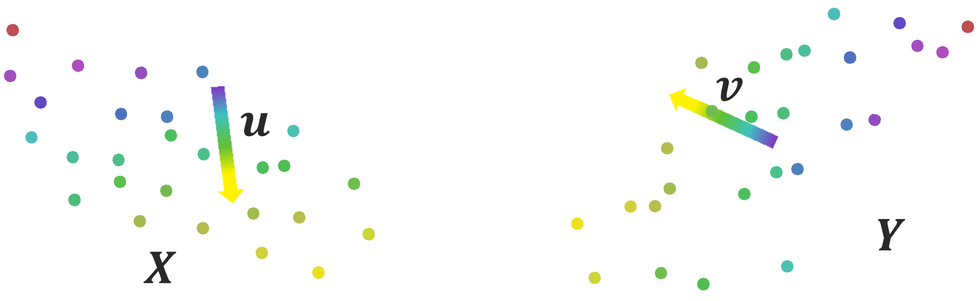
Ⓑ Coordinated representations:

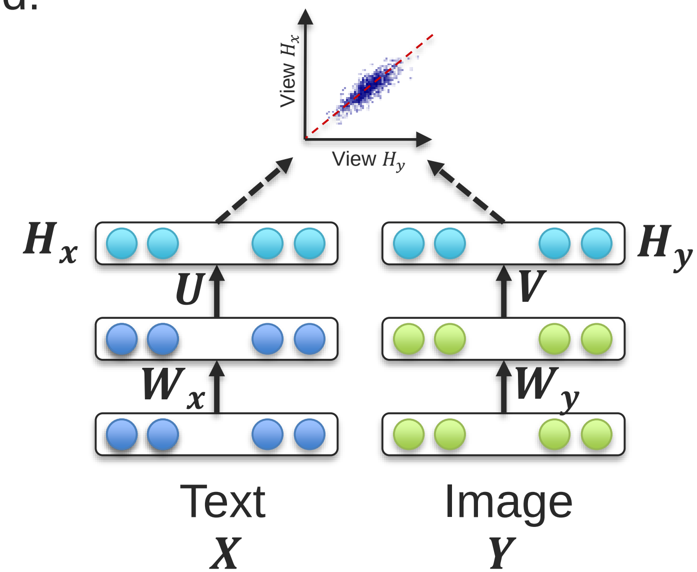


Coordinated Representation: Deep CCA

Learn linear projections that are maximally correlated:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr\left(\boldsymbol{u}^T \boldsymbol{X}, \boldsymbol{v}^T \boldsymbol{Y}\right)$$

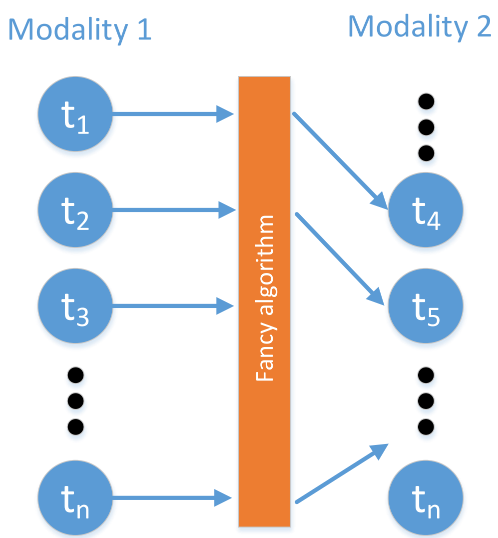




Andrew et al., ICML 2013

Core Challenge 2: Alignment

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



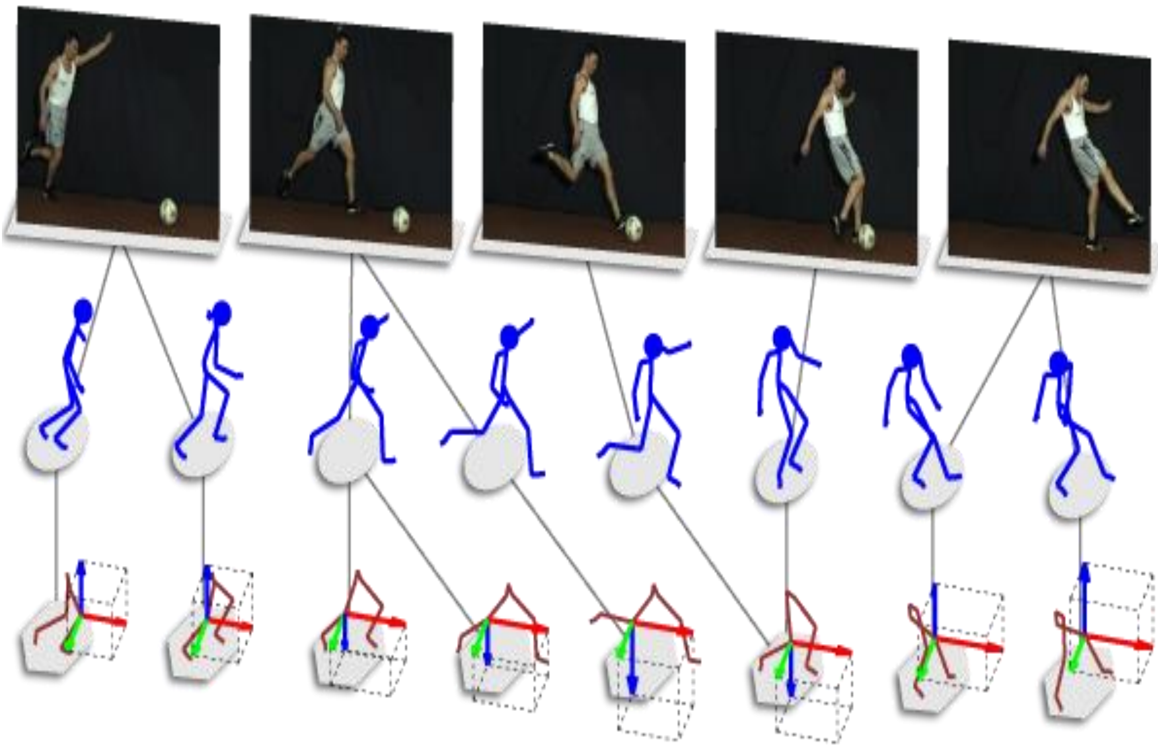
A **Explicit Alignment**

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

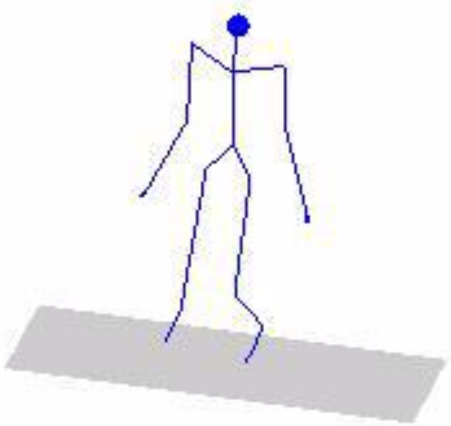
Temporal sequence alignment



Applications:
- Re-aligning asynchronous data
- Finding similar data across modalities (we can estimate the aligned cost)
- Event reconstruction from multiple sources

Alignment examples (multimodal)

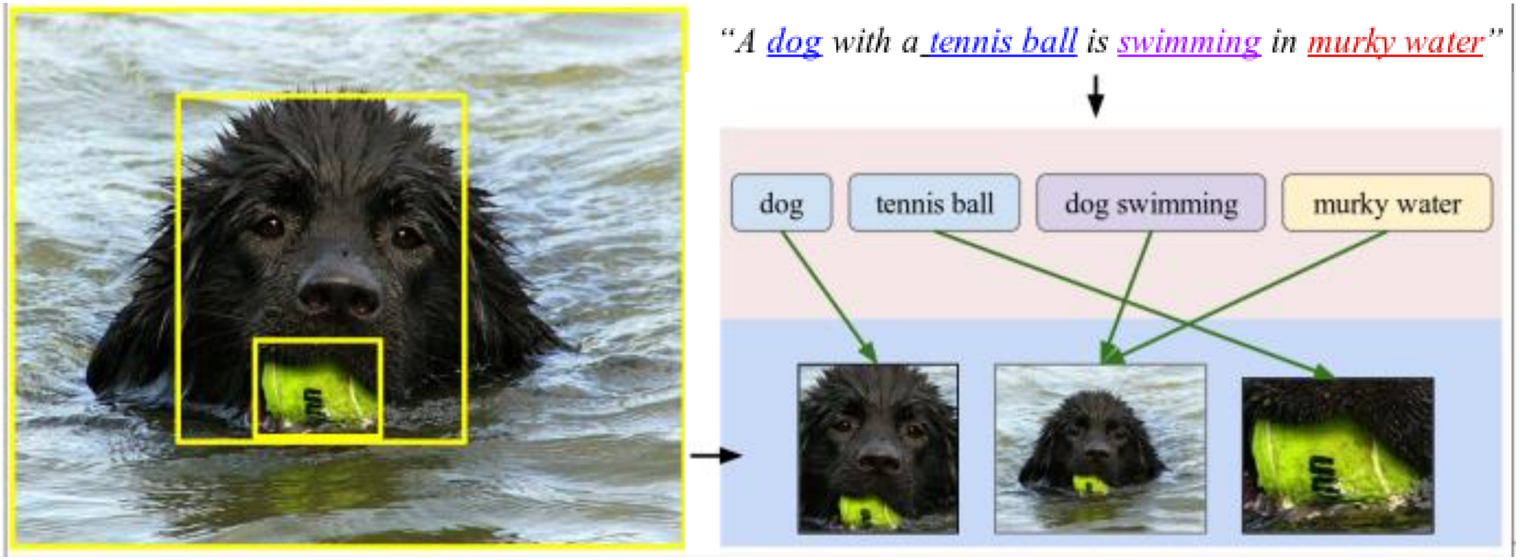
1/273



1/51

1/127

Implicit Alignment



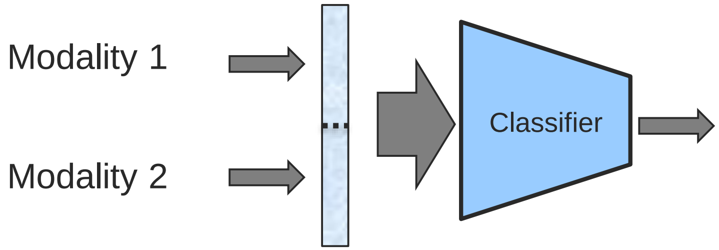
Karpathy et al., Deep Fragment Embeddings for Bidirectional Image Sentence Mapping, https://arxiv.org/pdf/1406.5679.pdf

Core Challenge 3: Fusion

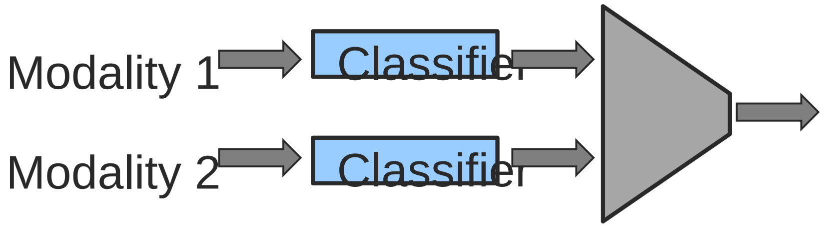
Definition: To join information from two or more modalities to perform a prediction task.

(A) Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion

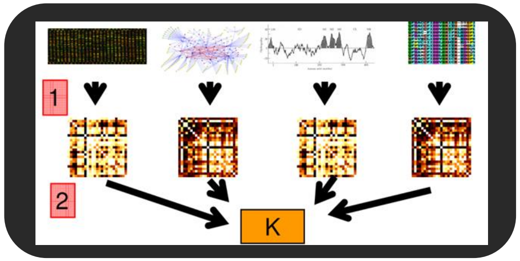


Core Challenge 3: Fusion

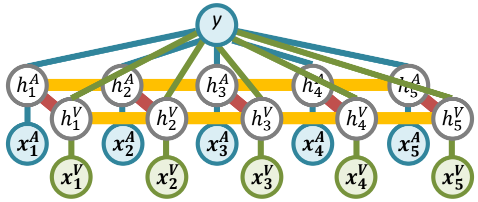
Definition: To join information from two or more modalities to perform a prediction task.

B Model-Based (Intermediate) Approaches

1) Deep neural networks
2) Kernel-based methods
3) Graphical models



Multiple kernel learning

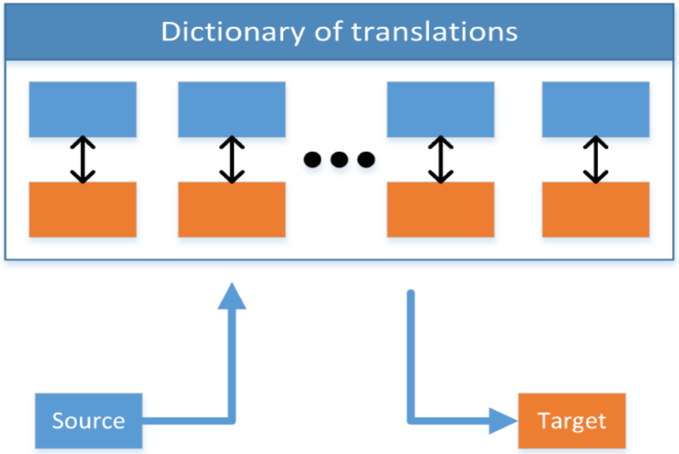


Multi-View Hidden CRF

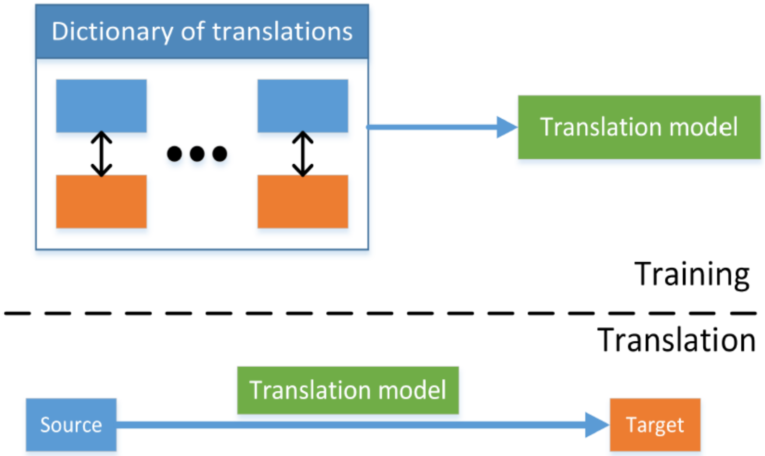
Core Challenge 4: Translation

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

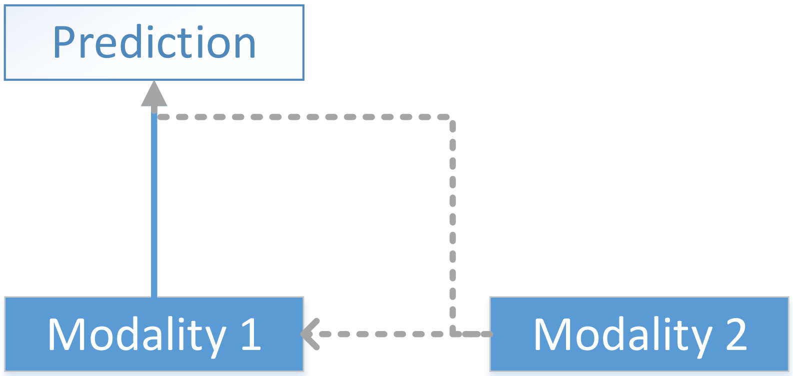
Visual gestures
(both speaker and listener gestures)

Transcriptions
+
Audio streams

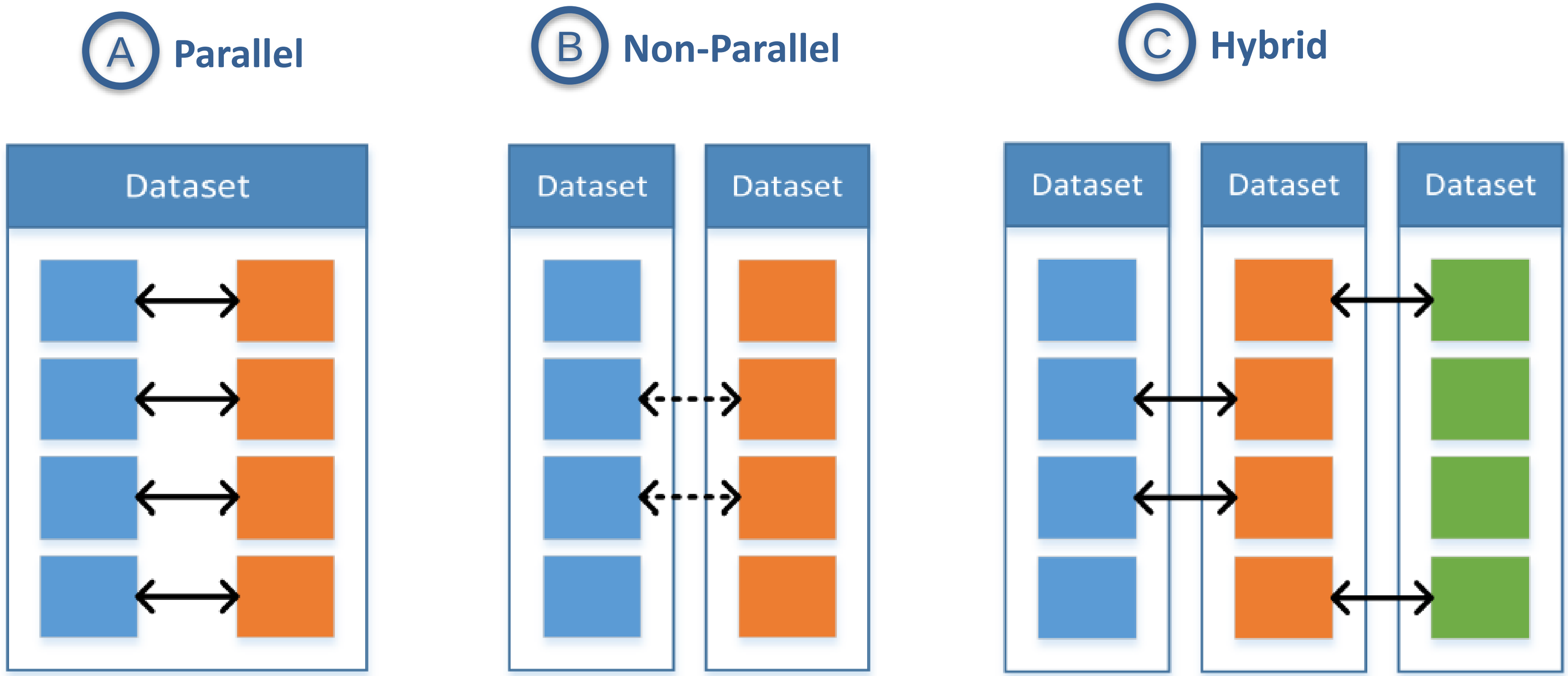
Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

Core Challenge 5: Co-Learning

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



Core Challenge 5: Co-Learning



Taxonomy of Multimodal Research [https://arxiv.org/abs/1705.09406]

Representation
- Joint
  - Neural networks
  - Graphical models
  - Sequential
- Coordinated
  - Similarity
  - Structured

Translation
- Example-based
  - Retrieval
  - Combination
- Model-based
  - Grammar-based
  - Encoder-decoder
  - Online prediction

Alignment
- Explicit
  - Unsupervised
  - Supervised
- Implicit
  - Graphical models
  - Neural networks

Fusion
- Model agnostic
  - Early fusion
  - Late fusion
  - Hybrid fusion
- Model-based
  - Kernel-based
  - Graphical models
  - Neural networks

Co-learning
- Parallel data
  - Co-training
  - Transfer learning
- Non-parallel data
  - Zero-shot learning
  - Concept grounding
  - Transfer learning
- Hybrid data
  - Bridging

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Multimodal Applications

[https://arxiv.org/abs/1705.09406]

	Challenges				
Applications	Representation	Translation	Fusion	Alignment	Co-learning
Speech Recognition and Synthesis					
Audio-visual Speech Recognition	✓		✓	✓	✓
(Visual) Speech Synthesis	✓	✓			
Event Detection					
Action Classification	✓		✓		✓
Multimedia Event Detection	✓		✓		✓
Emotion and Affect					
Recognition	✓		✓	✓	✓
Synthesis	✓	✓			
Media Description					
Image Description	✓	✓		✓	✓
Video Description	✓	✓	✓	✓	✓
Visual Question-Answering	✓		✓	✓	✓
Media Summarization	✓	✓	✓		
Multimedia Retrieval					
Cross Modal retrieval	✓	✓		✓	✓
Cross Modal hashing	✓				✓

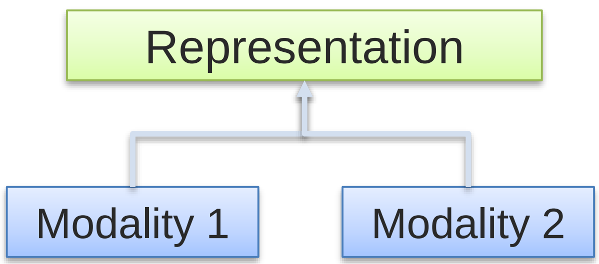
Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Multimodal Representations

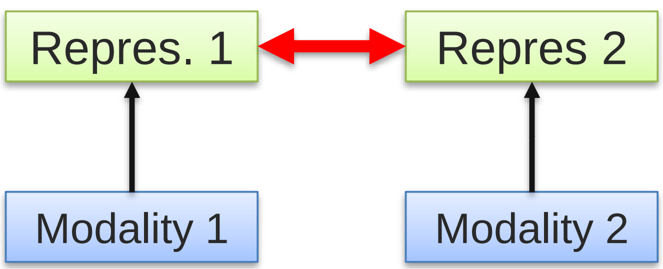
Core Challenge: Representation

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

Ⓐ Joint representations:

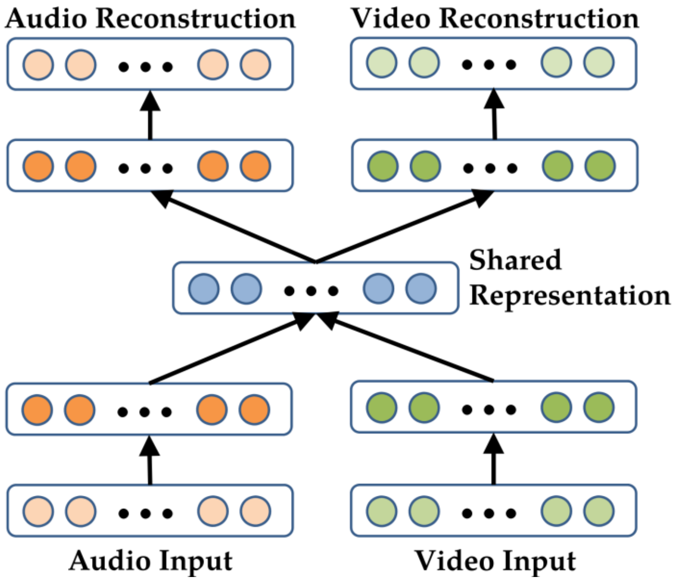


Ⓑ Coordinated representations:



Deep Multimodal autoencoders

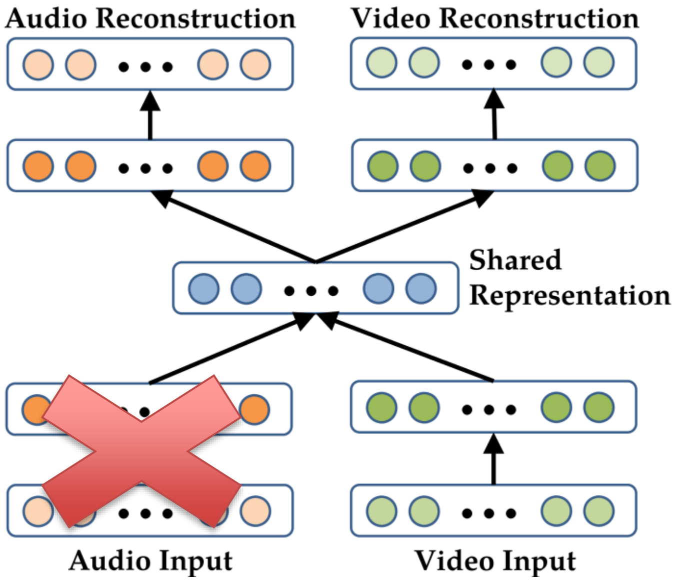
- A deep representation learning approach
- A bimodal auto-encoder
  - Used for Audio-visual speech recognition



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

Deep Multimodal autoencoders - training

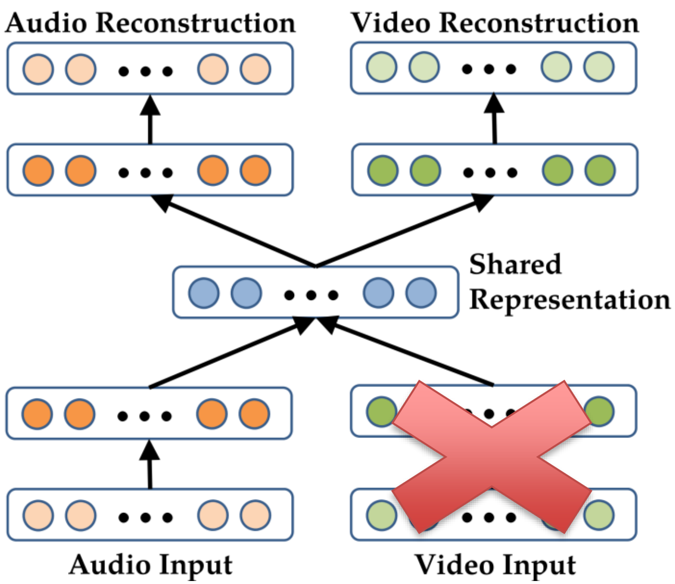
- Individual modalities can be pretrained
  - RBMs
  - Denoising Autoencoders
- To train the model to reconstruct the other modality
  - Use both
  - Remove audio



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

Deep Multimodal autoencoders - training

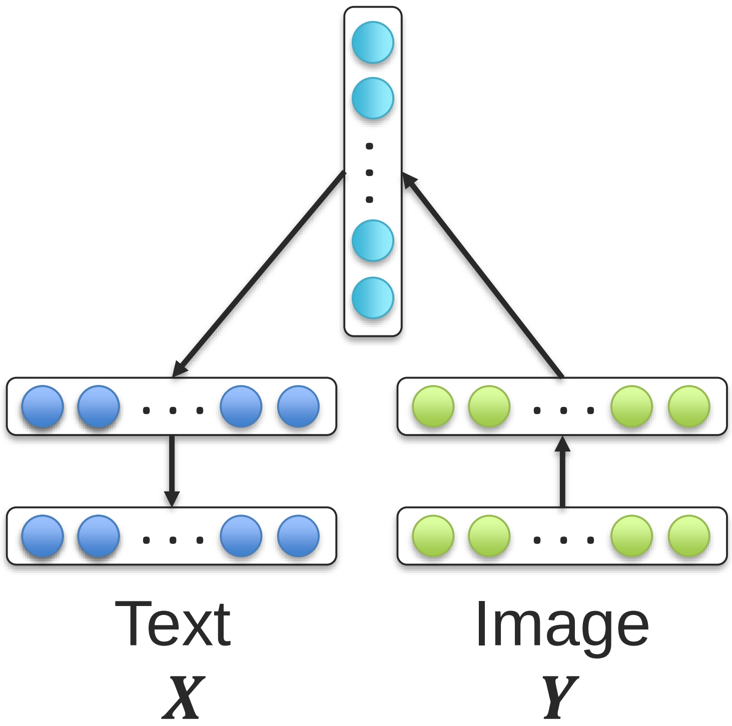
- Individual modalities can be pretrained
    - RBMs
    - Denoising Autoencoders
- To train the model to reconstruct the other modality
    - Use both
    - Remove audio
    - Remove video



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

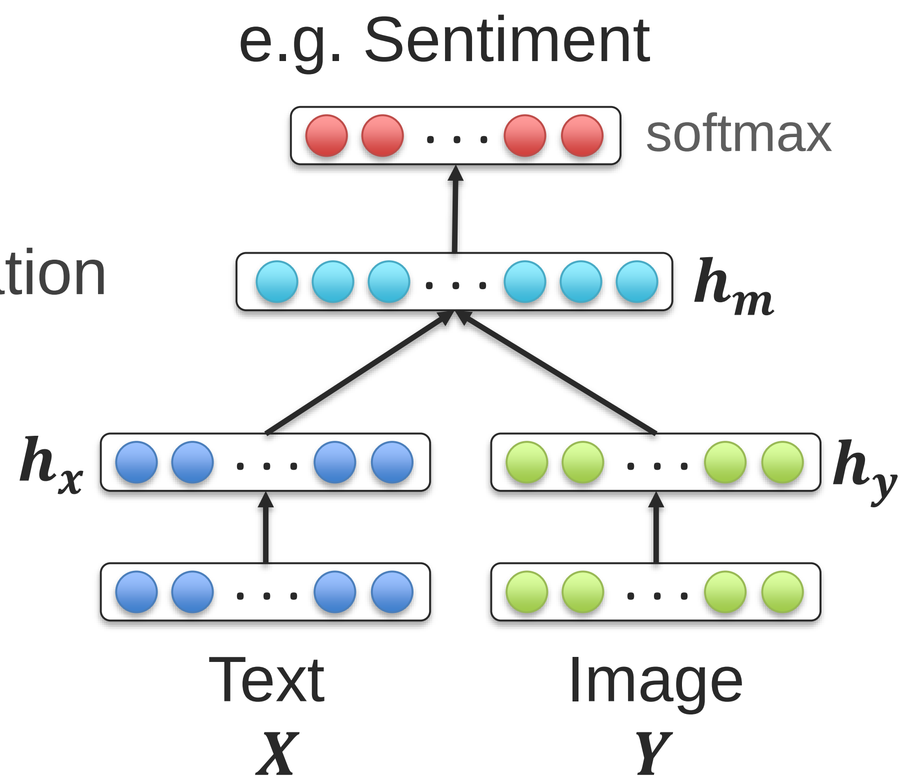
Multimodal Encoder-Decoder

- 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?



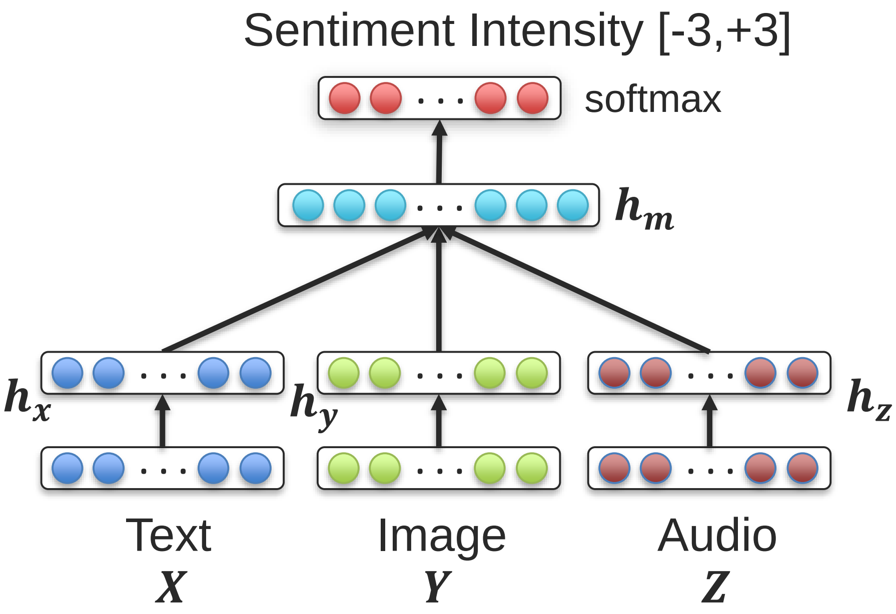
Multimodal Sentiment Analysis

MOSI dataset (Zadeh et al, 2016)

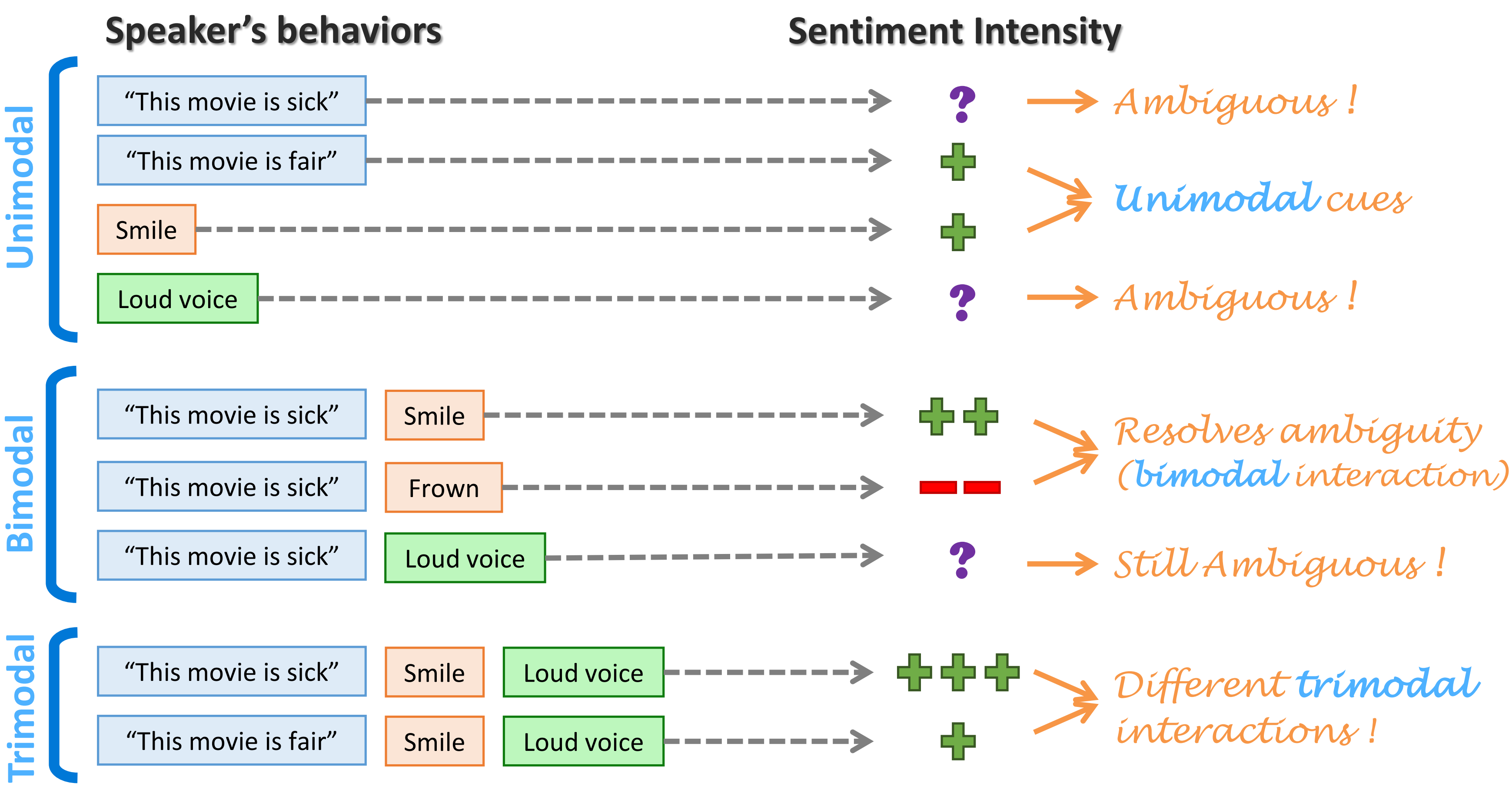
- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

Multimodal joint representation:

$$h_m = f\left(W \cdot [h_x, h_y, h_z]\right)$$



Unimodal, Bimodal and Trimodal Interactions

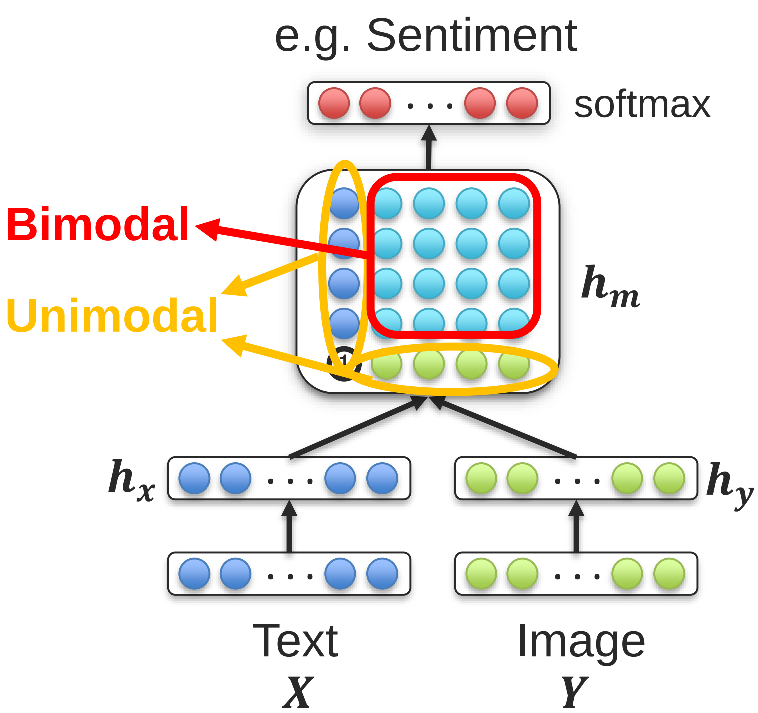


Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

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

Important !

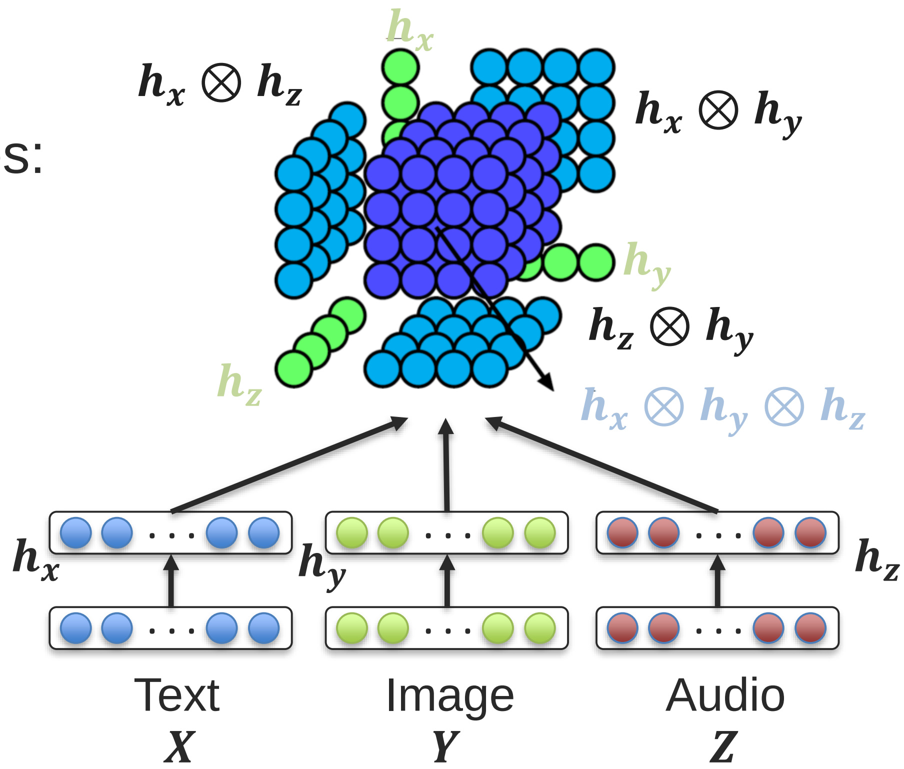


[Zadeh, Jones and Morency, **EMNLP 2017**]

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**]

Experimental Results – MOSI Dataset

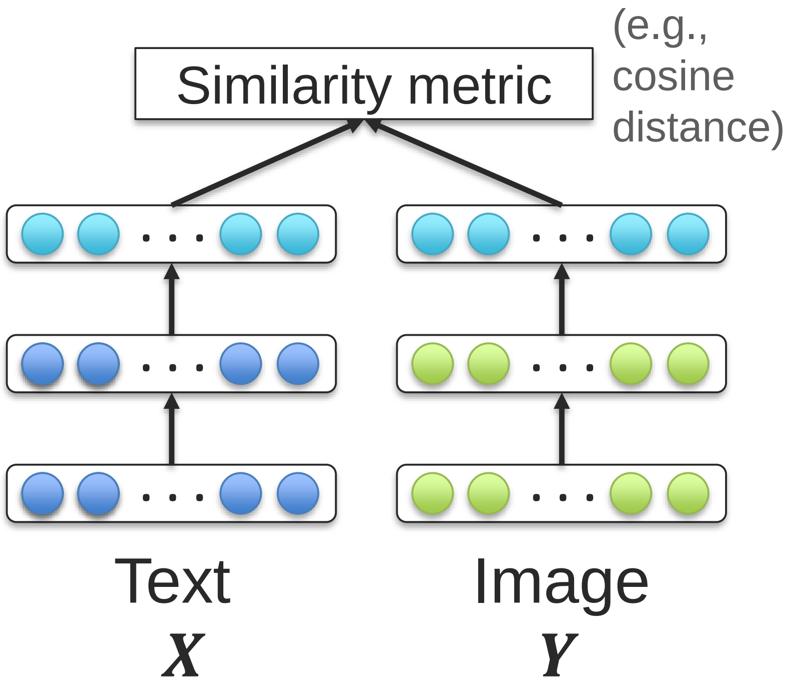
Multimodal Baseline	Binary Acc(%)	Binary F1	5-class Acc(%)	Regression MAE	Regression r
Random	50.2	48.7	23.9	1.88	-
C-MKL	73.1	75.2	35.3	-	-
SAL-CNN	73.0	-	-	-	-
SVM-MD	71.6	72.3	32.0	1.10	0.53
RF	71.4	72.1	31.9	1.11	0.51
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
Human	85.7	87.5	53.9	0.71	0.82
Δ^{SOTA}	↑ 4.0	↑ 2.7	↑ 6.7	↓ 0.23	↑ 0.17

Improvement over State-Of-The-Art

Baseline	Binary Acc(%)	Binary F1	5-class Acc(%)	Regression MAE	Regression r
TFN$_{language}$	74.8	75.6	38.5	0.99	0.61
TFN$_{visual}$	66.8	70.4	30.4	1.13	0.48
TFN$_{acoustic}$	65.1	67.3	27.5	1.23	0.36
TFN$_{bimodal}$	75.2	76.0	39.6	0.92	0.65
TFN$_{trimodal}$	74.5	75.0	38.9	0.93	0.65
TFN$_{notrimodal}$	75.3	76.2	39.7	0.919	0.66
TFN	**77.1**	**77.9**	**42.0**	**0.87**	**0.70**
TFN$_{early}$	75.2	76.2	39.0	0.96	0.63

Coordinated Multimodal Representations

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

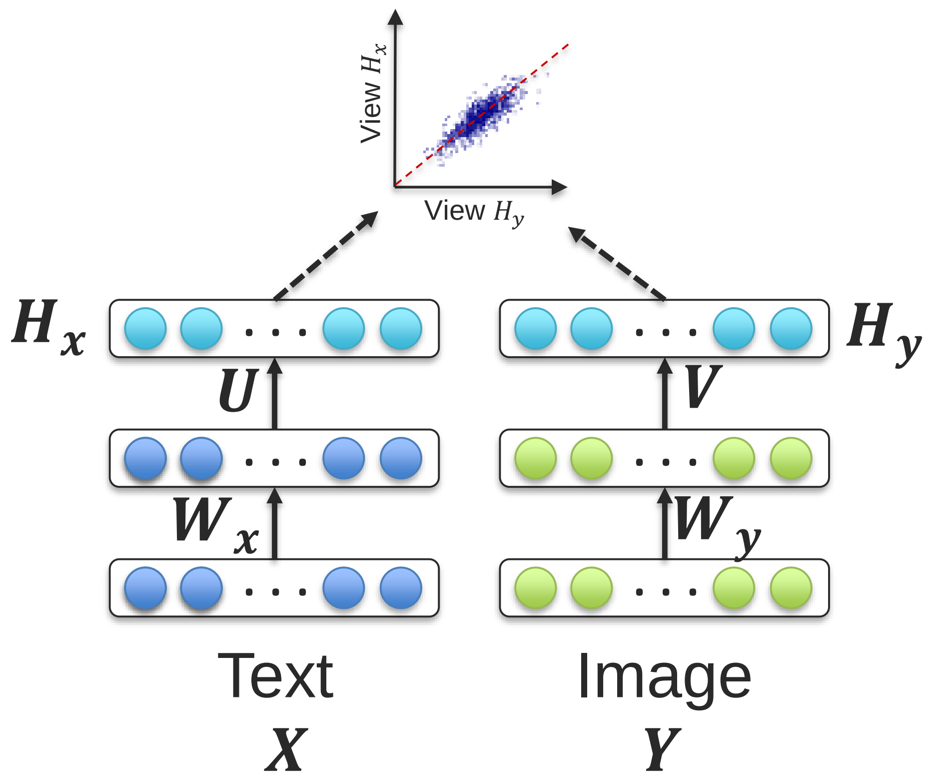


Deep Canonical Correlation Analysis

Same objective function as CCA:

$$\underset{V,U,W_x,W_y}{\operatorname{argmax}}\ corr(\boldsymbol{H_x}, \boldsymbol{H_y})$$

1. Linear projections maximizing correlation
2. Orthogonal projections
3. Unit variance of the projection vectors

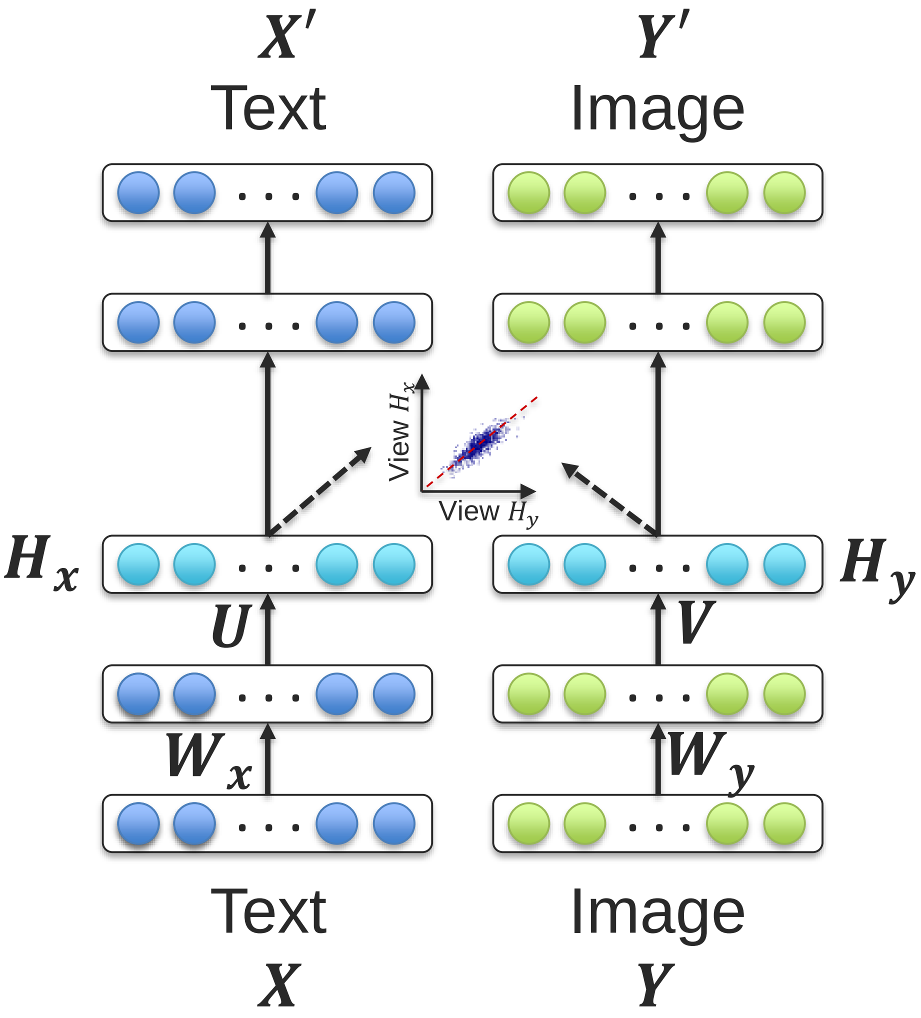


Andrew et al., ICML 2013

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

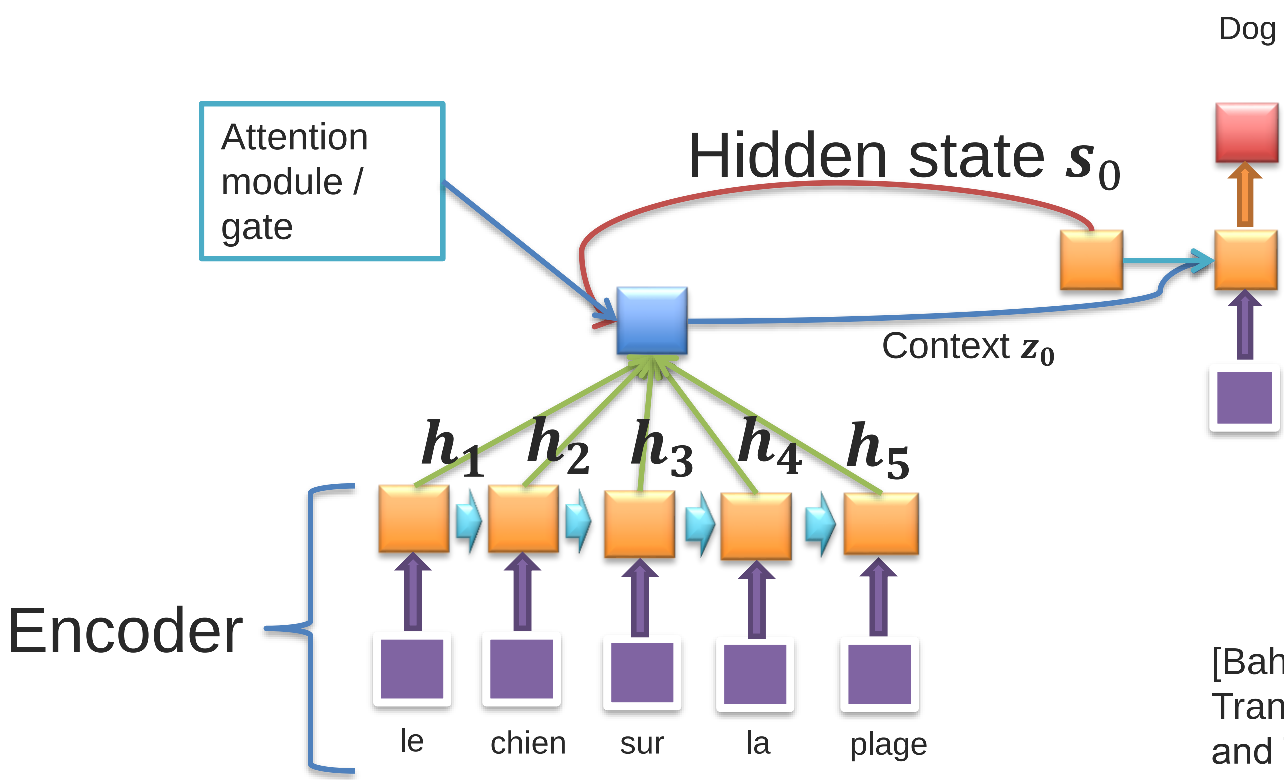
Implicit alignment

Machine Translation

- Given a sentence in one language translate it to another

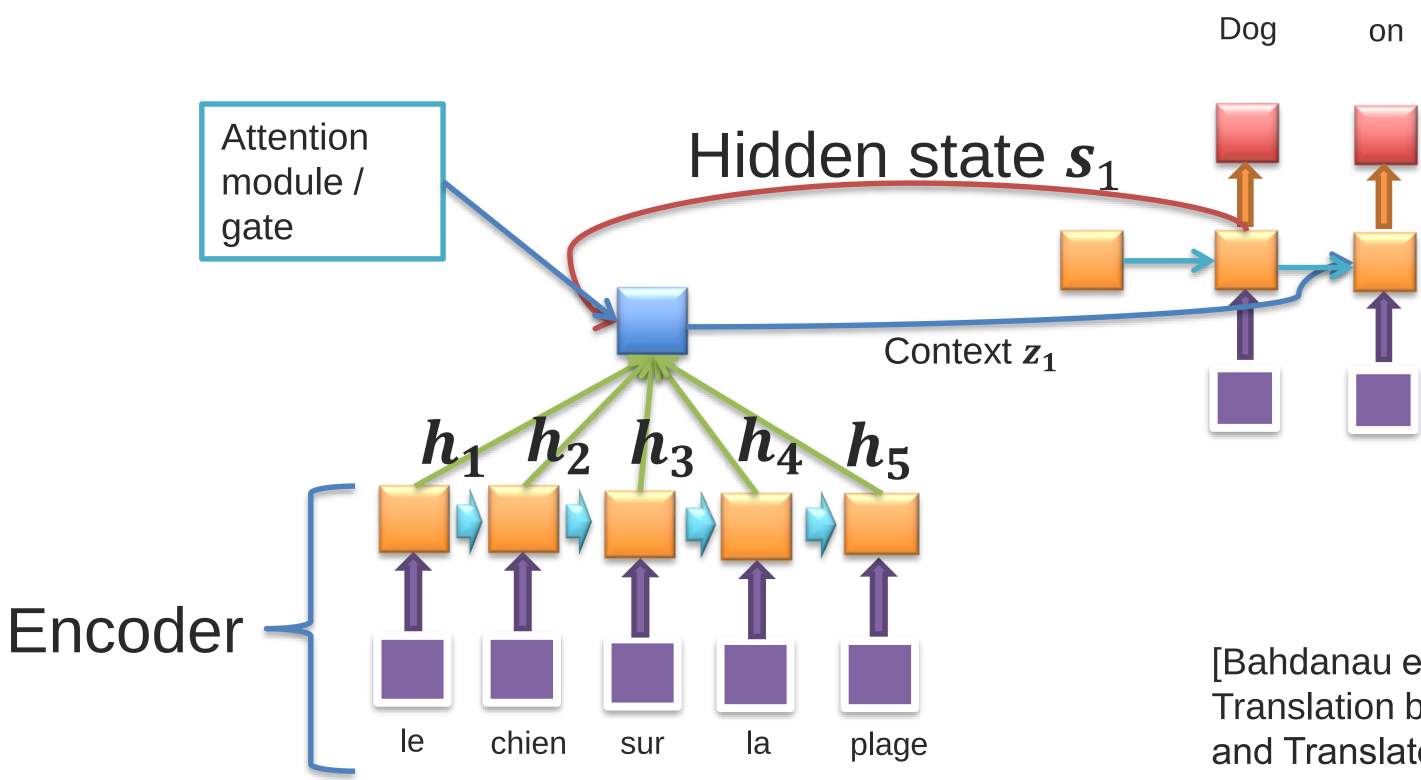
Dog on the beach → le chien sur la plage

Attention Model for Machine Translation



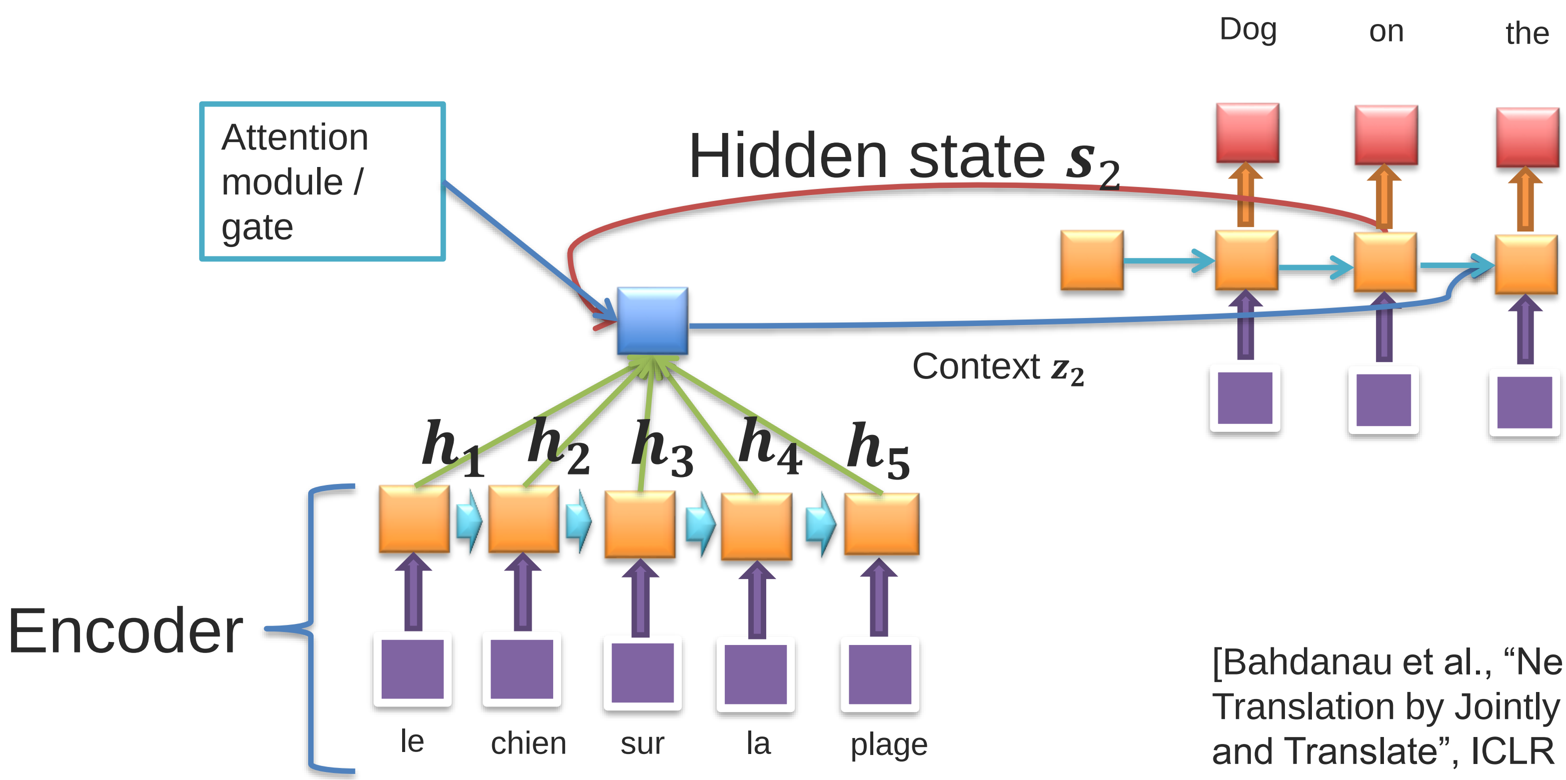
[Bahdanau et al., “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015]

Attention Model for Machine Translation



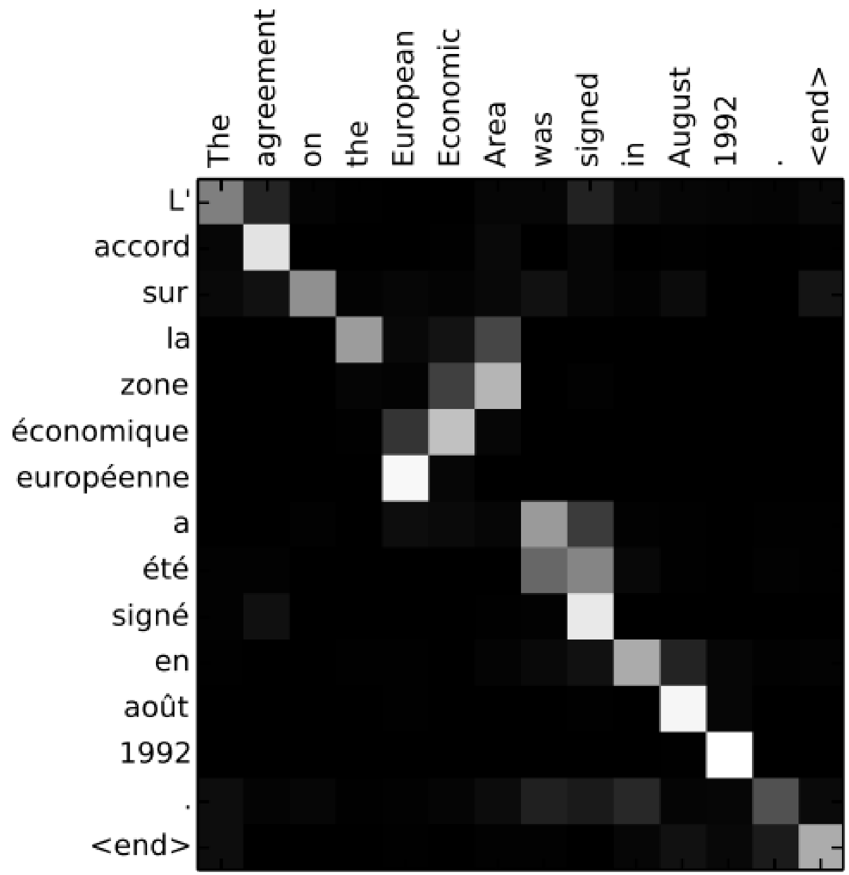
[Bahdanau et al., “Neural Machine Translation by Jointly Learning to Align and Translate”, ICLR 2015]

Attention Model for Machine Translation

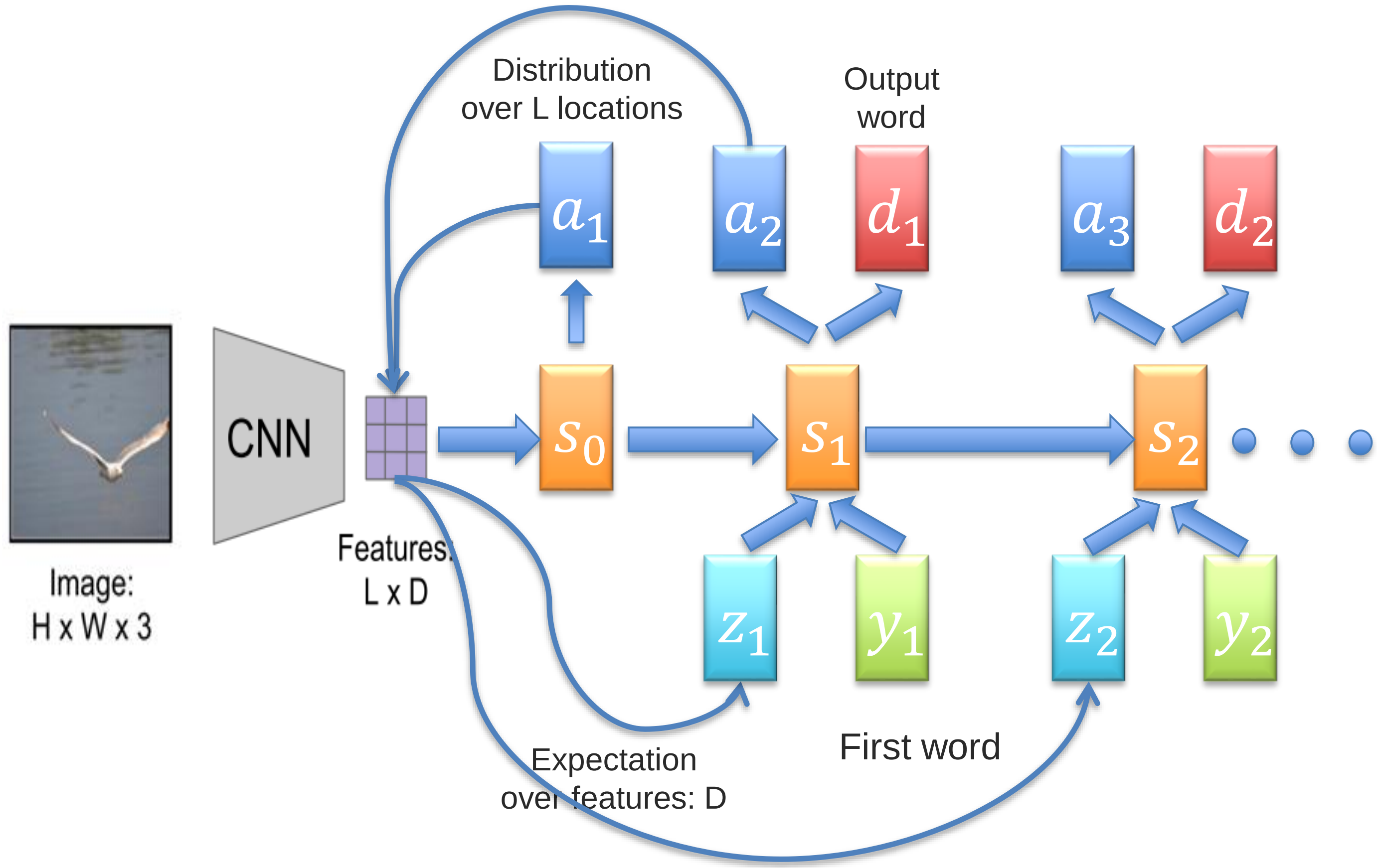


[Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015]

Attention Model for Machine Translation



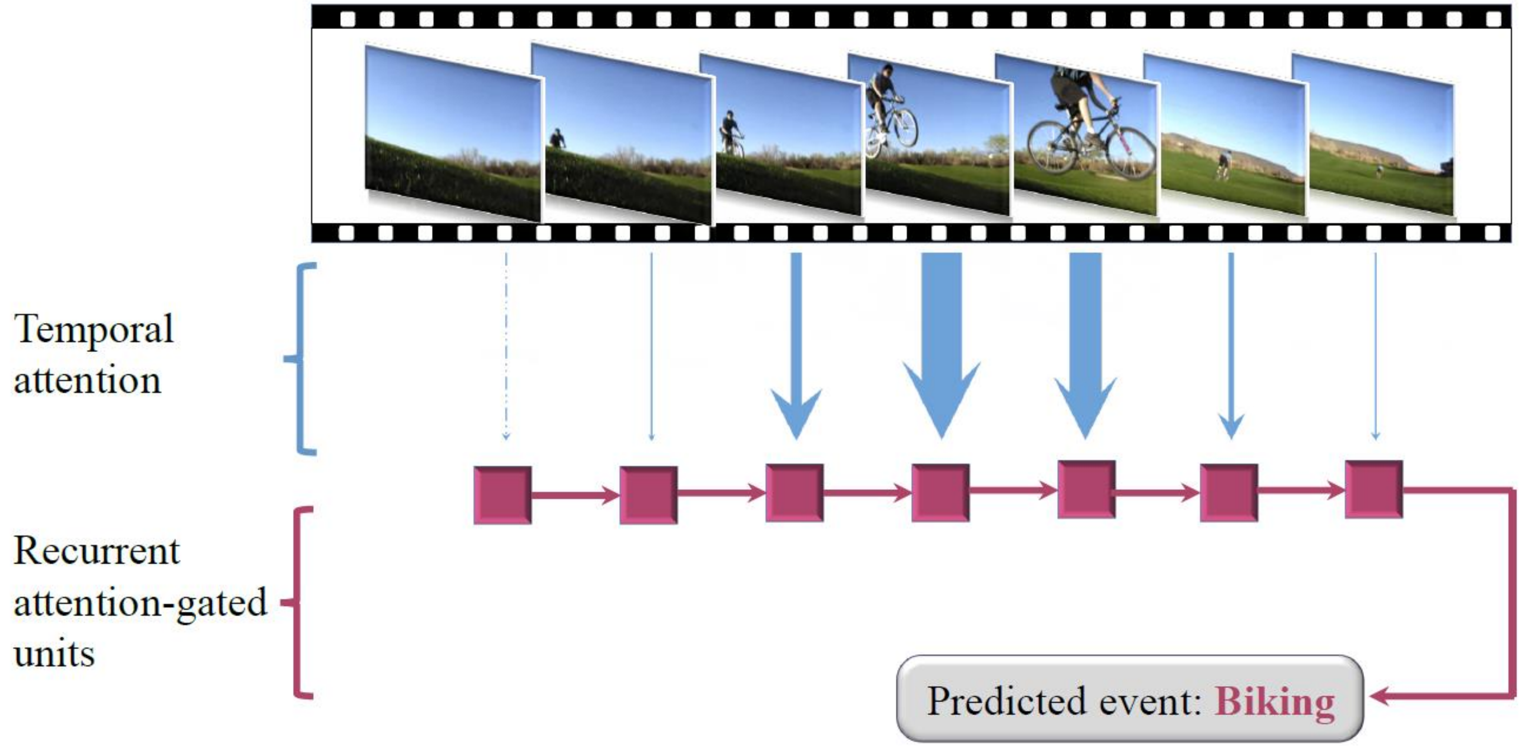
Attention Model for Image Captioning



Attention Model for Image Captioning

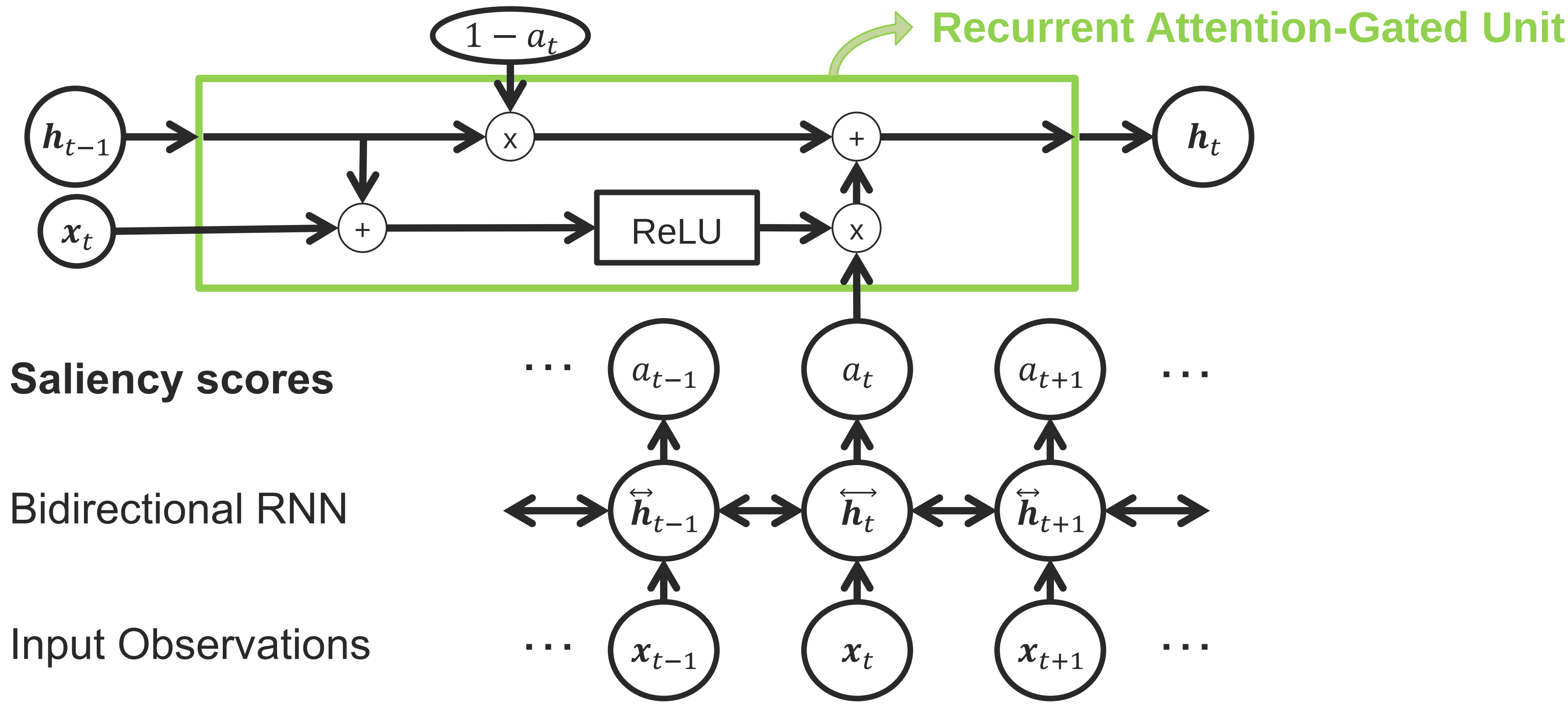
Xu et.al., ICML 2015

Attention Model for Video Sequences



[Pei, Baltrušaitis, Tax and Morency. Temporal Attention-Gated Model for Robust Sequence Classification, **CVPR**, *2017]*

Temporal Attention-Gated Model (TAGM)



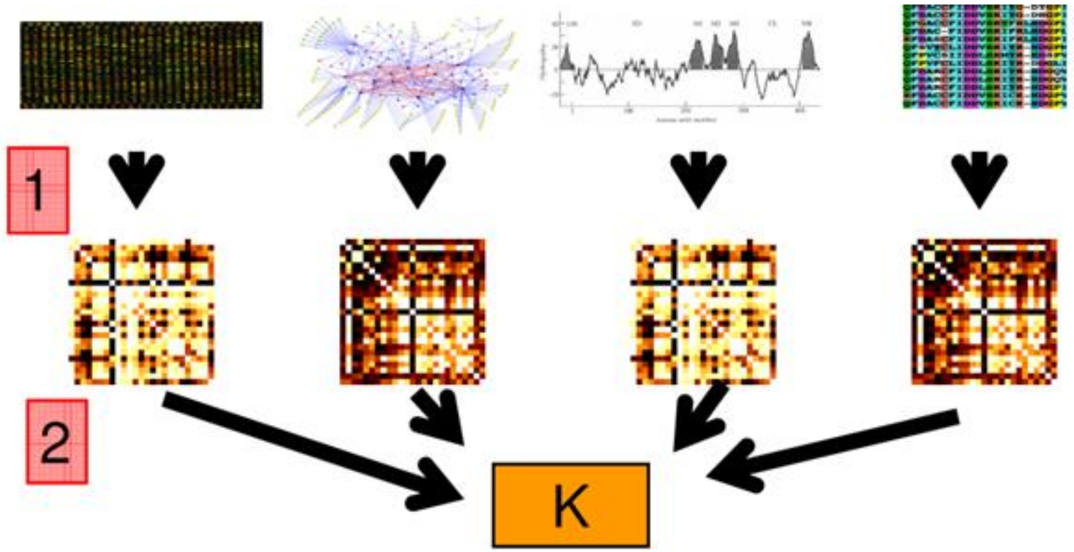
Temporal Attention Gated Model (TAGM)

[Pei, Baltrušaitis, Tax and Morency. Temporal Attention-Gated Model for Robust Sequence Classification, **CVPR**, *2017*]

Multimodal Fusion

Multiple Kernel Learning

- Pick a family of kernels for each modality and learn which kernels are important for the classification case
- Generalizes the idea of Support Vector Machines
- Works as well for unimodal and multimodal data, very little adaptation is needed



[Lanckriet 2004]

Multimodal Fusion for Sequential Data

Modality-*private* structure

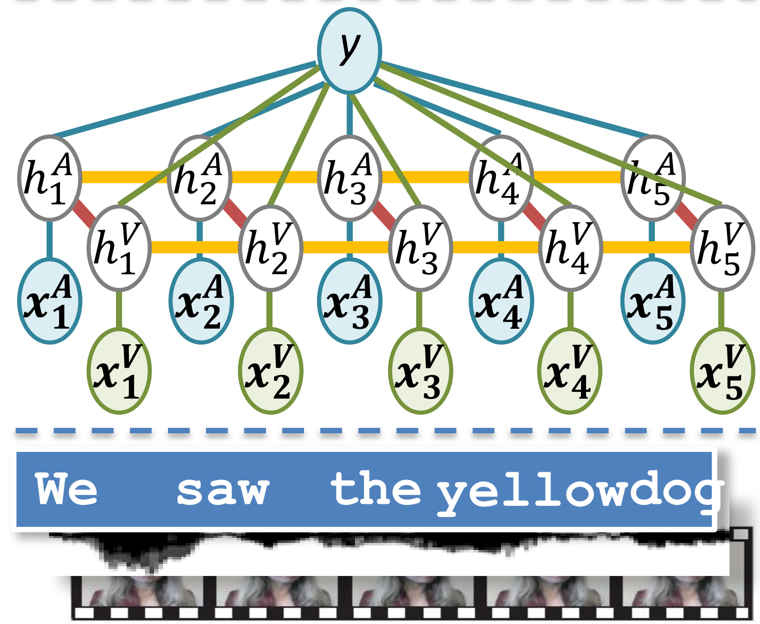
- Internal grouping of observations

Modality-*shared* structure

- Interaction and synchrony

$$p(y|\,\boldsymbol{x}^{\boldsymbol{A}}, \boldsymbol{x}^{V}; \boldsymbol{\theta}) = \sum_{\boldsymbol{h}^{A}, \boldsymbol{h}^{V}} p\left(y, \boldsymbol{h}^{\boldsymbol{A}}, \boldsymbol{h}^{V} \middle| \boldsymbol{x}^{A}, \boldsymbol{x}^{V}; \boldsymbol{\theta}\right)$$

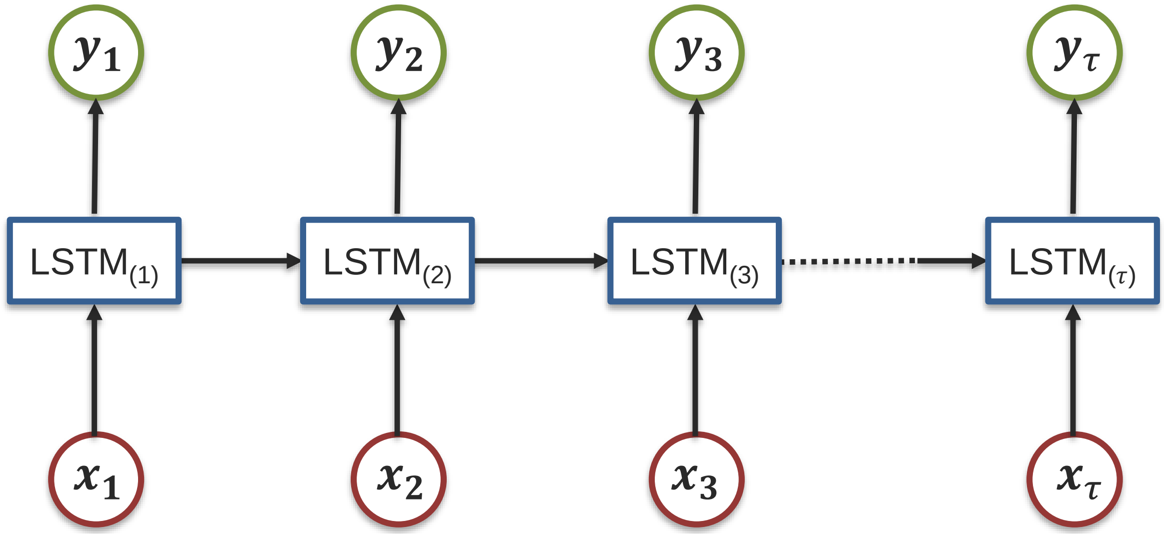
Multi-View
Hidden Conditional Random Field



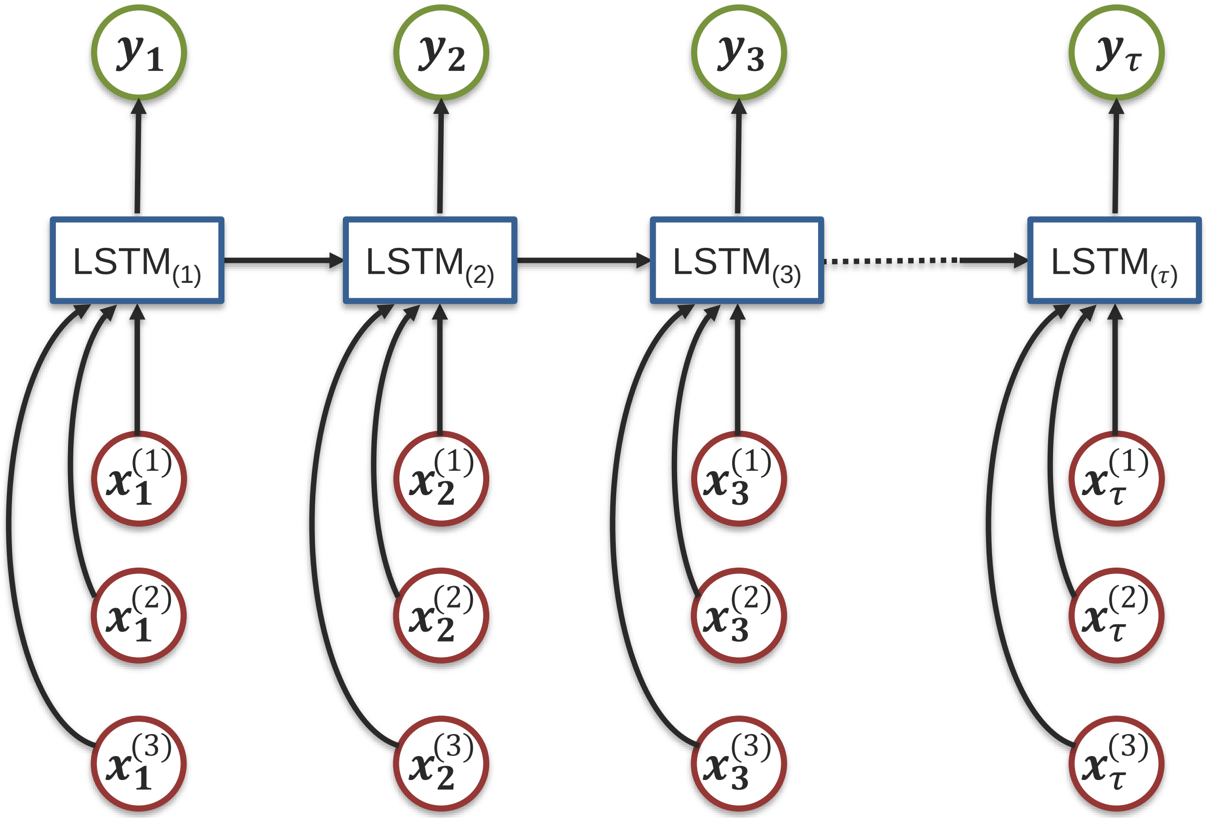
➢ Approximate inference using loopy-belief

[Song, Morency and Davis, CVPR 2012]

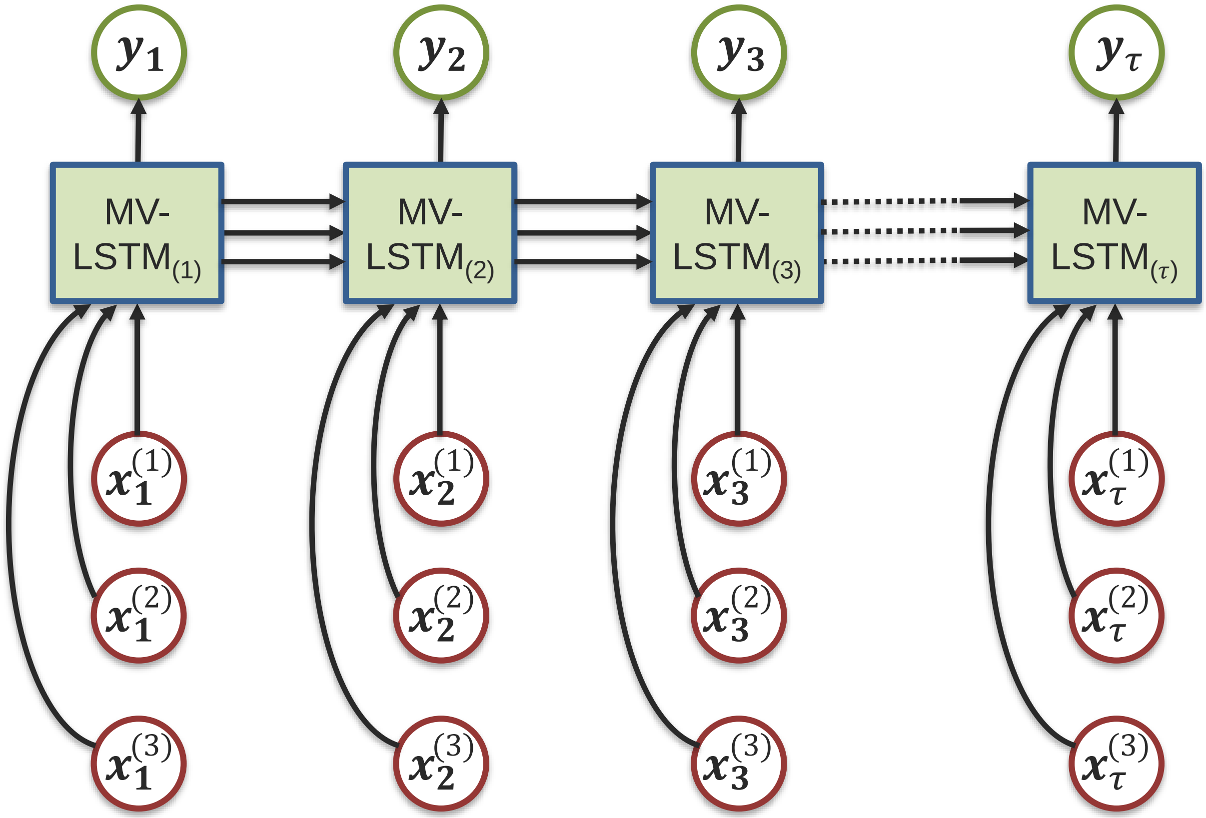
Sequence Modeling with LSTM



Multimodal Sequence Modeling – Early Fusion

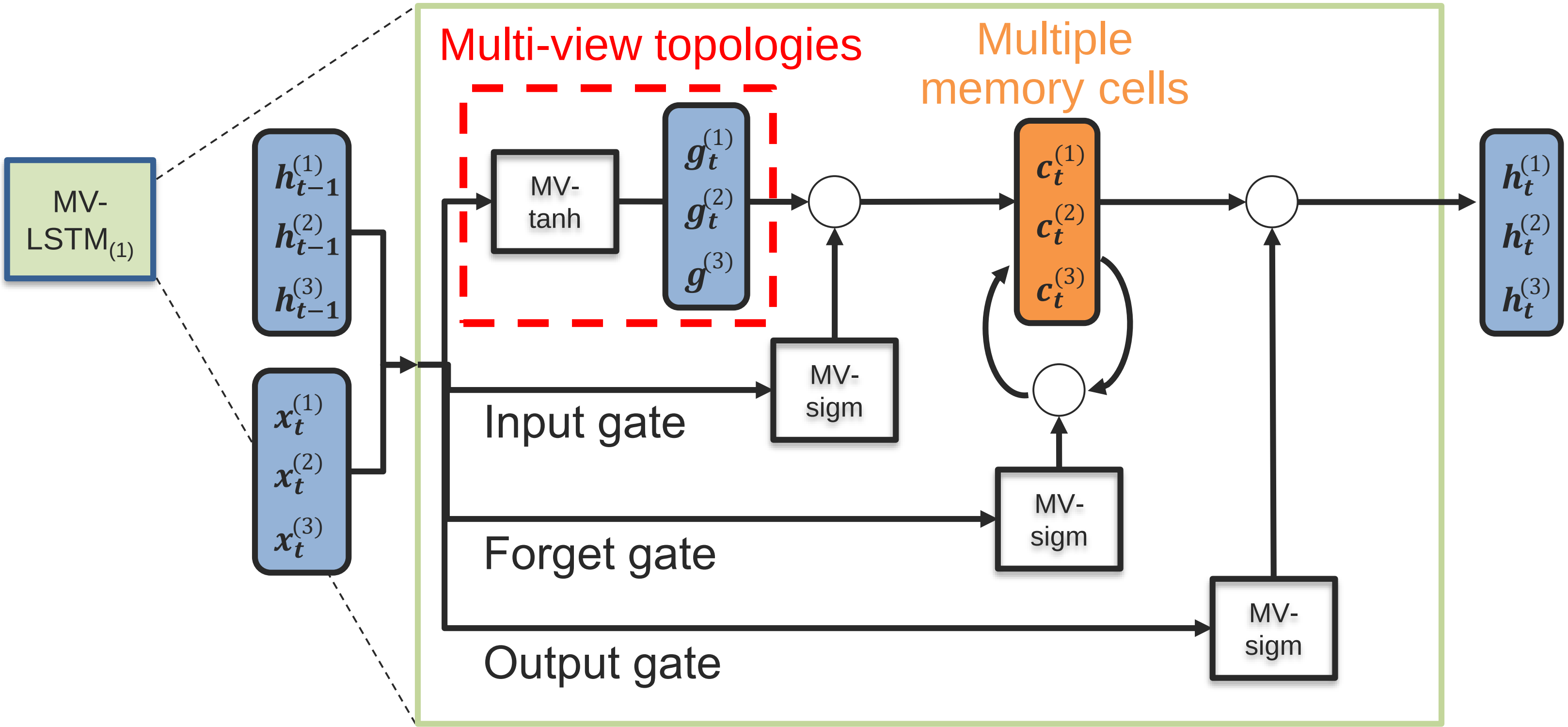


Multi-View Long Short-Term Memory (MV-LSTM)



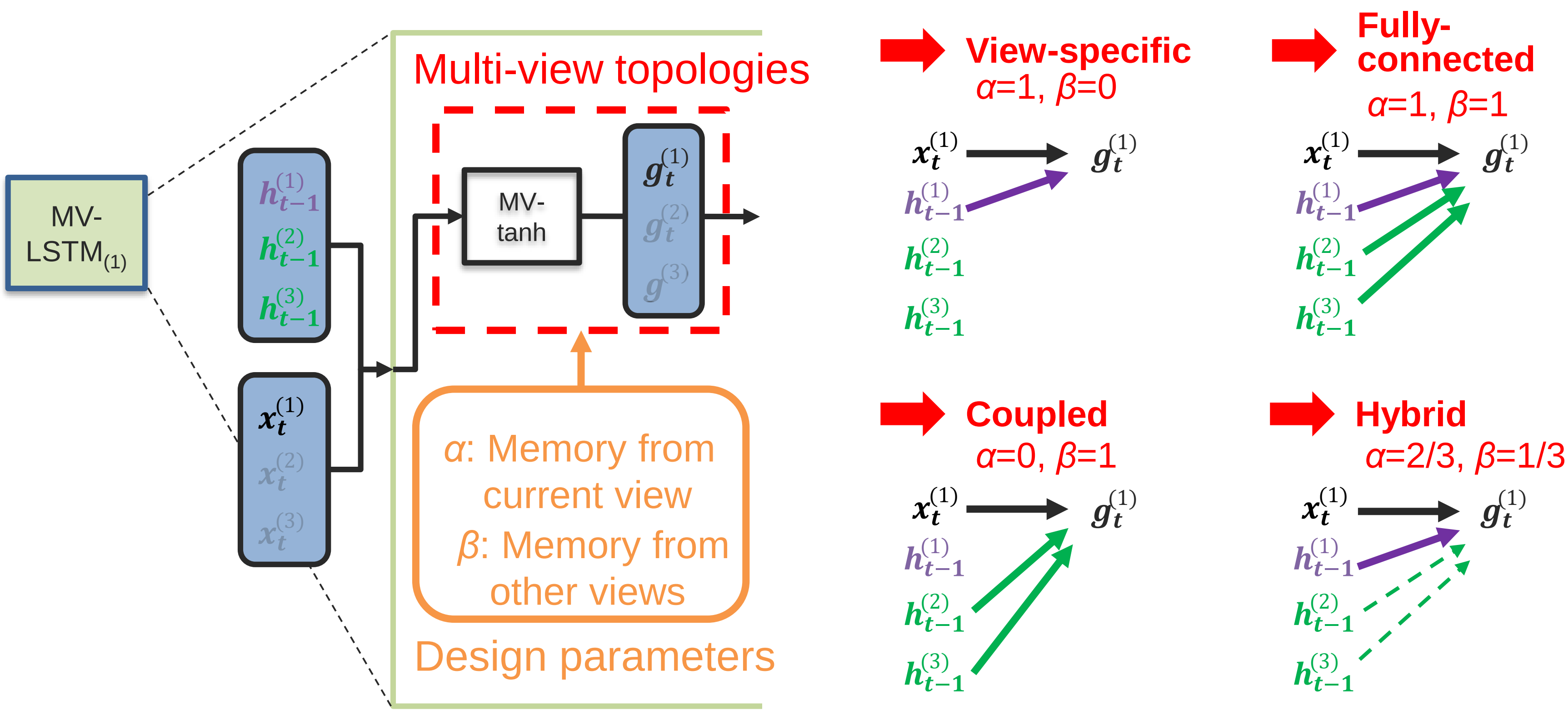
[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

Multi-View Long Short-Term Memory



[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

Topologies for Multi-View LSTM



[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

Multi-View Long Short-Term Memory (MV-LSTM)

Multimodal prediction of children engagement

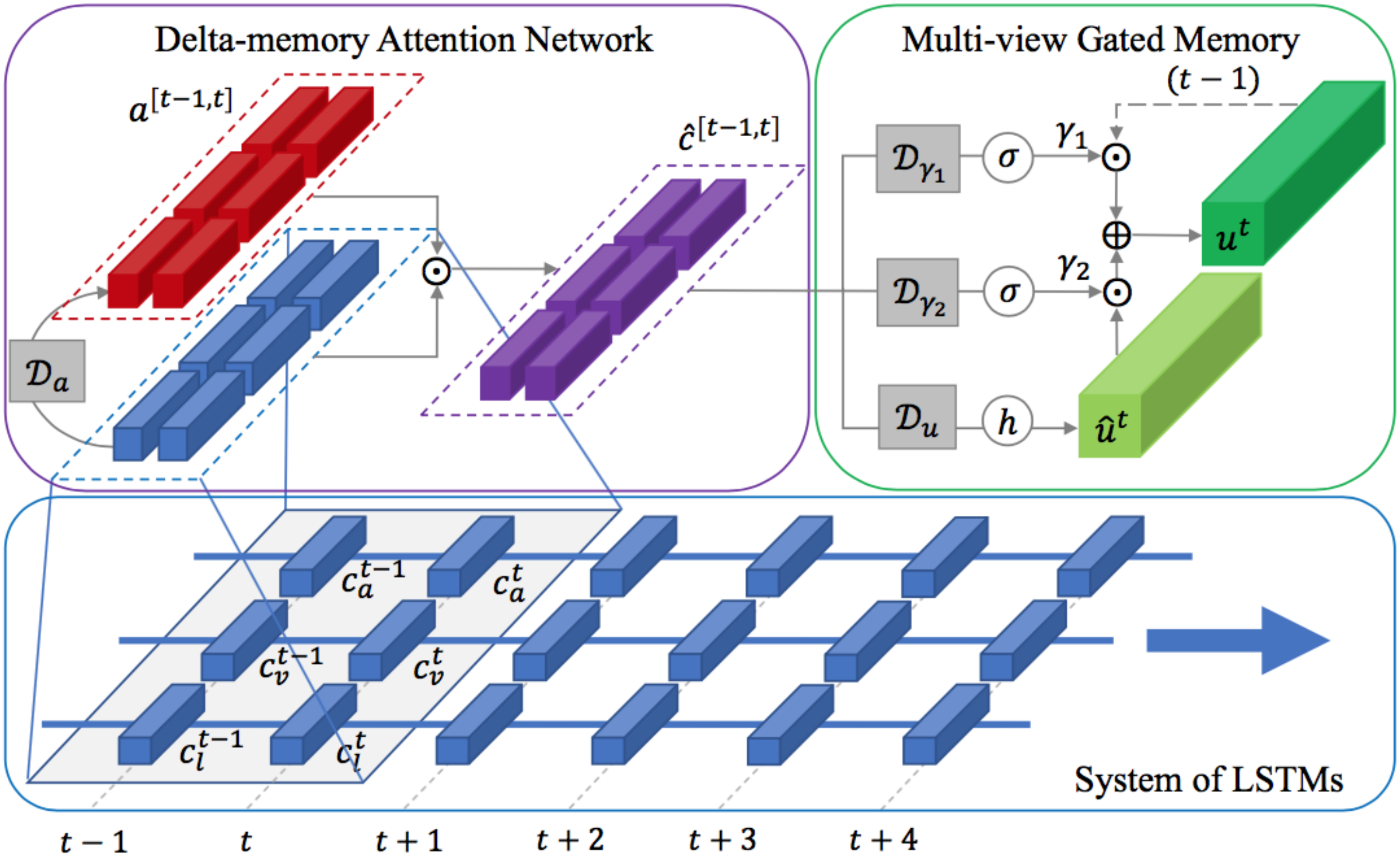
Class labels	Model	Precision	Recall	F1
Easy to engage	LSTM (Early fusion)	0.75	0.81	0.78
	MV-LSTM Full	0.81	0.81	0.81
	MV-LSTM Coupled	0.79	0.81	0.80
	MV-LSTM Hybrid	**0.80**	**0.86**	**0.83**
Difficult to engage	LSTM (Early fusion)	0.63	0.55	0.59
	MV-LSTM Full	0.68	0.68	0.68
	MV-LSTM Coupled	0.67	0.64	0.65
	MV-LSTM Hybrid	**0.74**	**0.64**	**0.68**

[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

Memory Based

- A memory accumulates multimodal information over time.
- From the representations throughout a source network.
- No need to modify the structure of the source network, only attached the memory.

Memory Based



[Zadeh et al., Memory Fusion Network for Multi-view Sequential Learning, AAAI 2018]

Multimodal Machine Learning

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**