Microsoft Research
Faculty Summit
2017

Multimodal Machine Learning

(or Deep Learning for Multimodal Systems)

Louis-Philippe Morency
Carnegie Mellon University

Integrative AI Systems

Robots

Virtual Humans

Ubiquitous

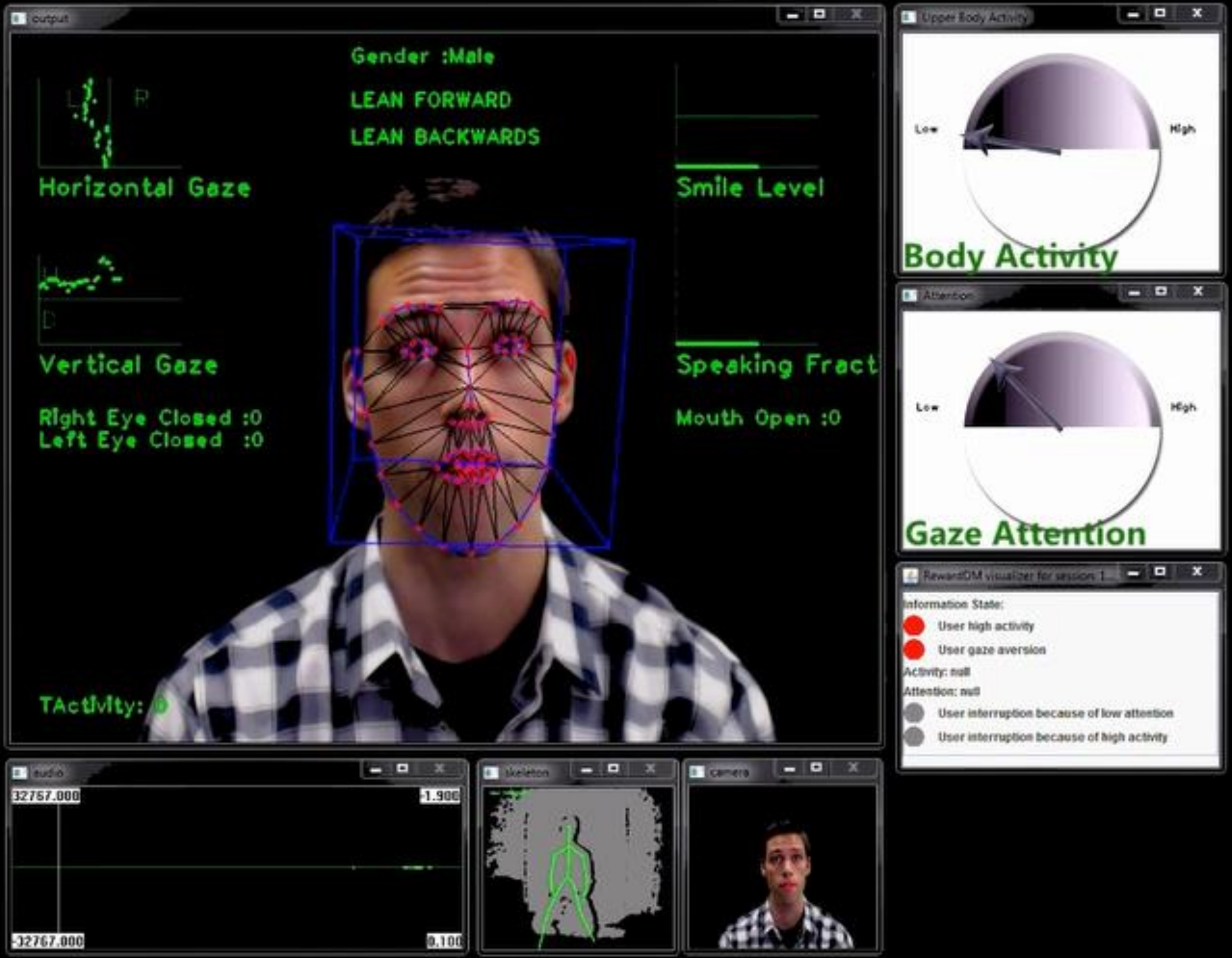
Mobile

Online

skype YouTube

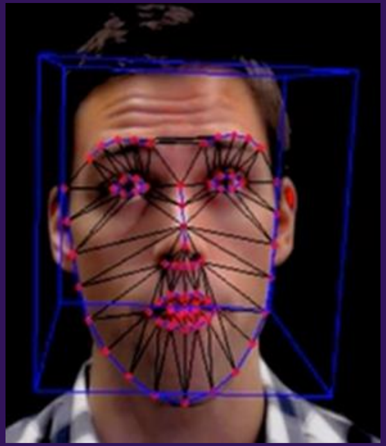
Wearable

MultiSense



SimSensei

Human Multimodal Behaviors



Verbal

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

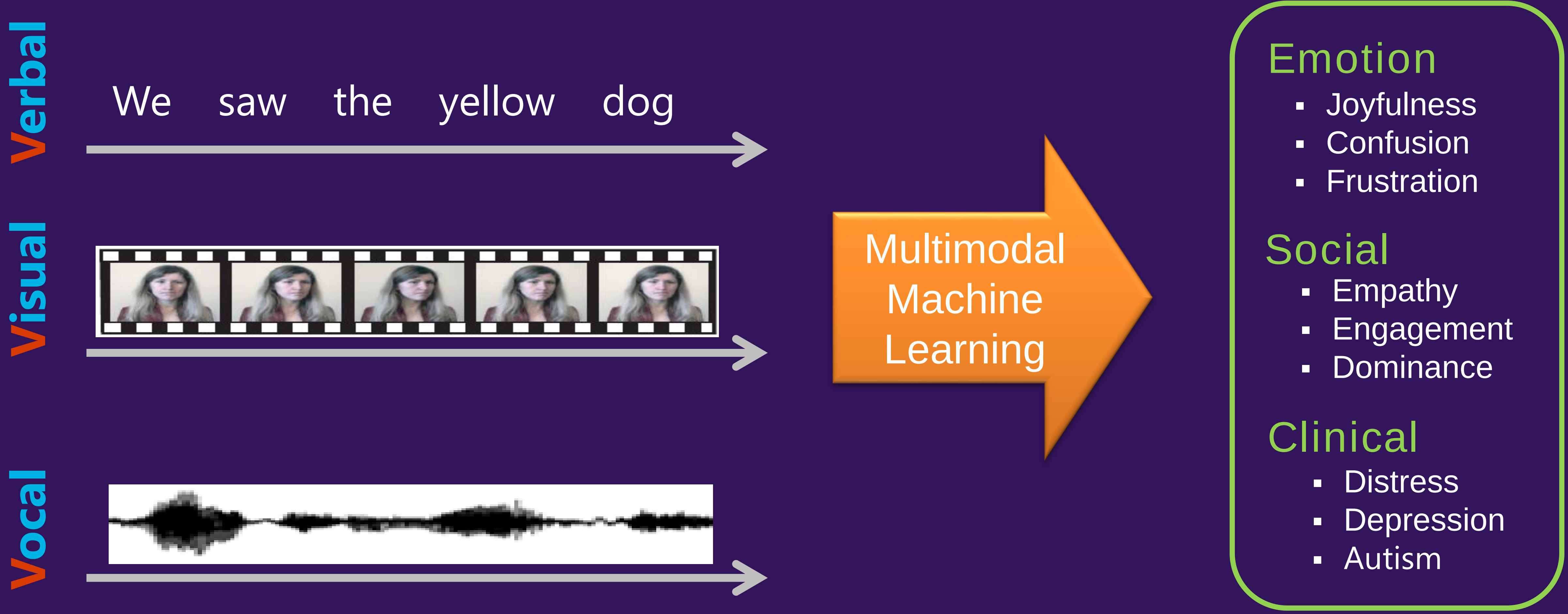
Vocal

- **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

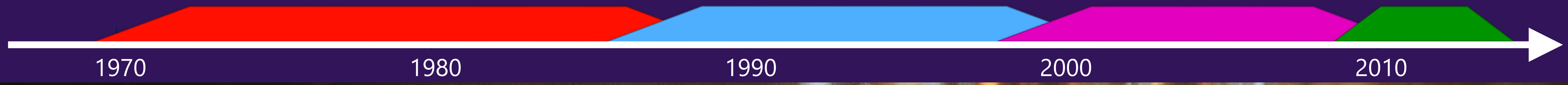
Multimodal Machine Learning



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 presentation



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

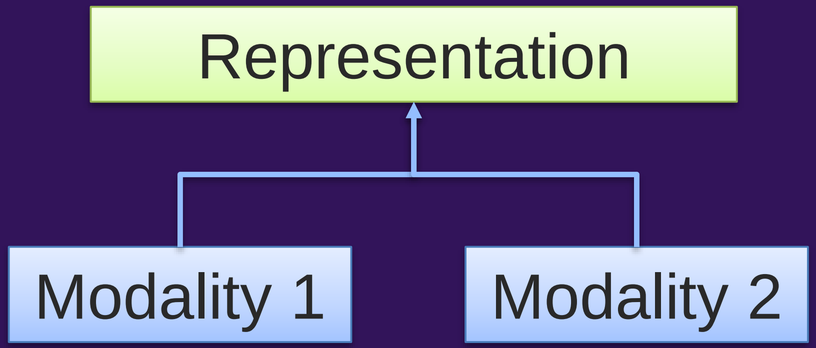
https://arxiv.org/abs/1705.09406

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

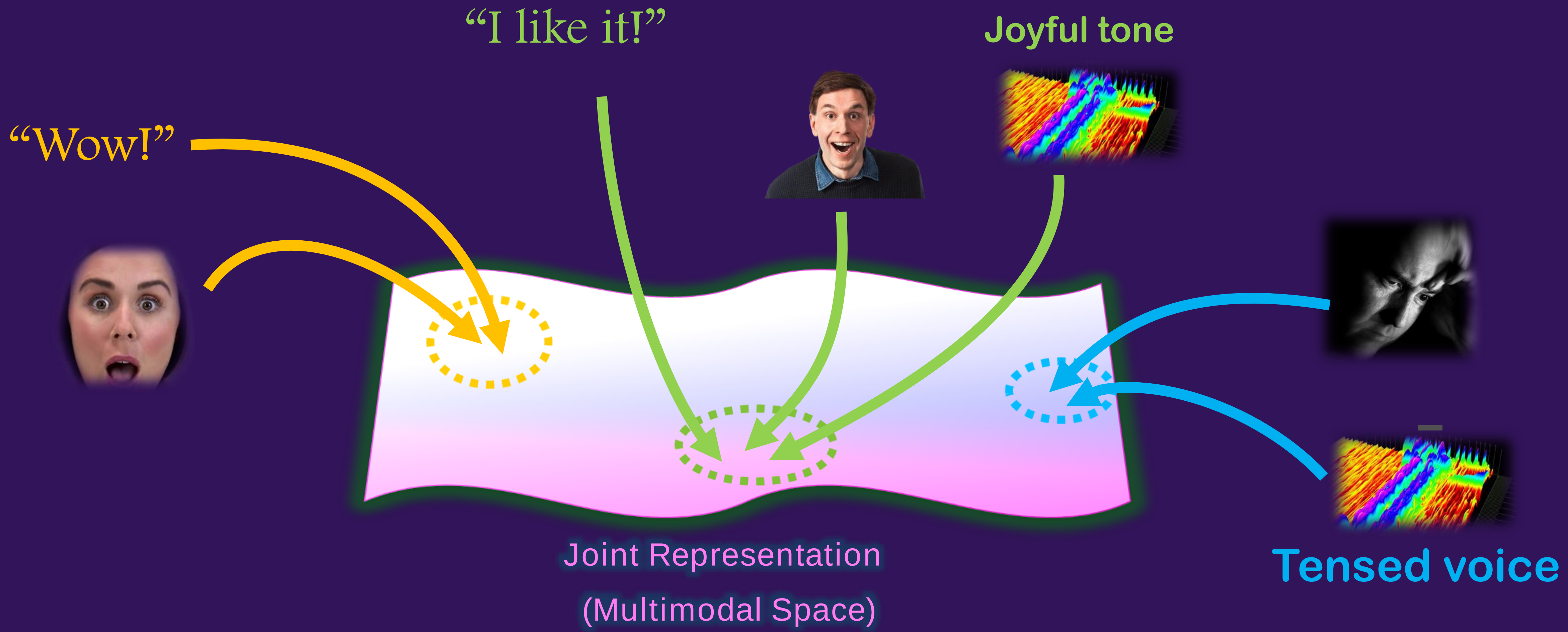
Core Challenge 1: Representation

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

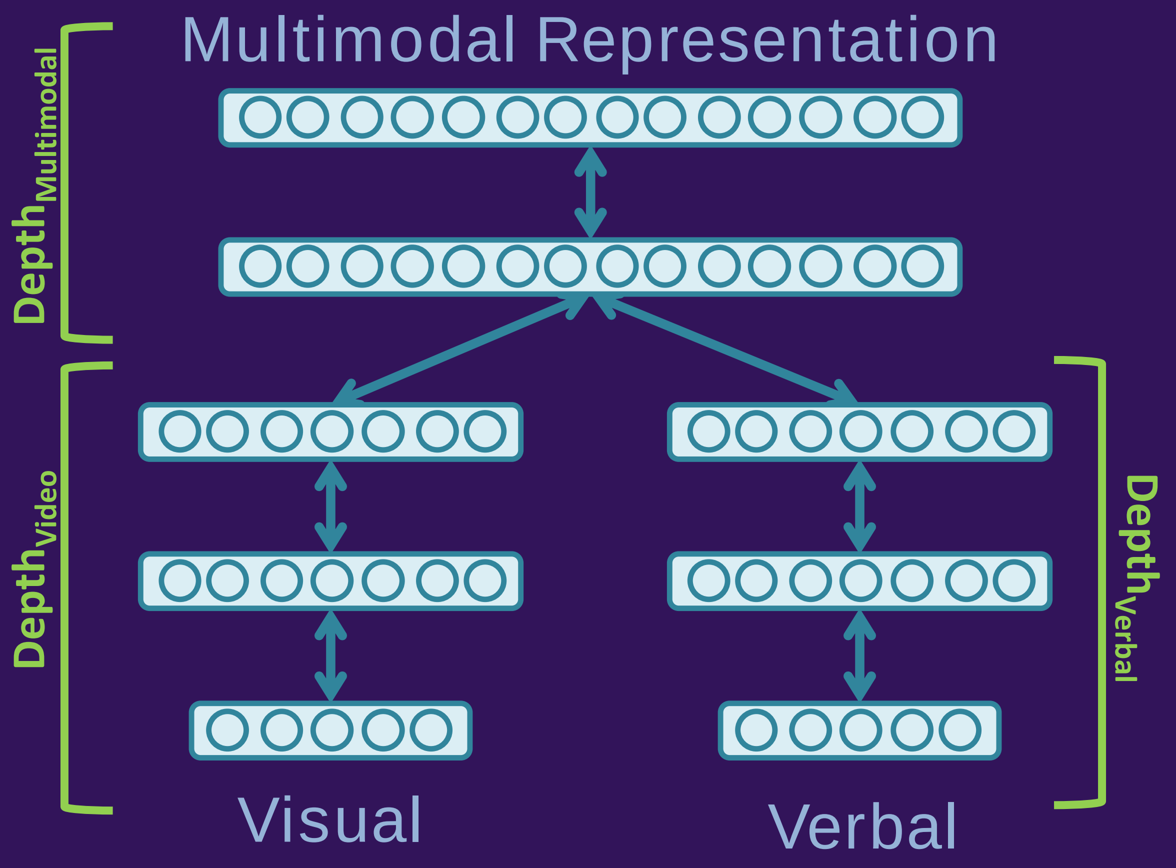
[Srivastava and Salahutdinov, NIPS 2012]

- Multimodal Deep Boltzmann Machine

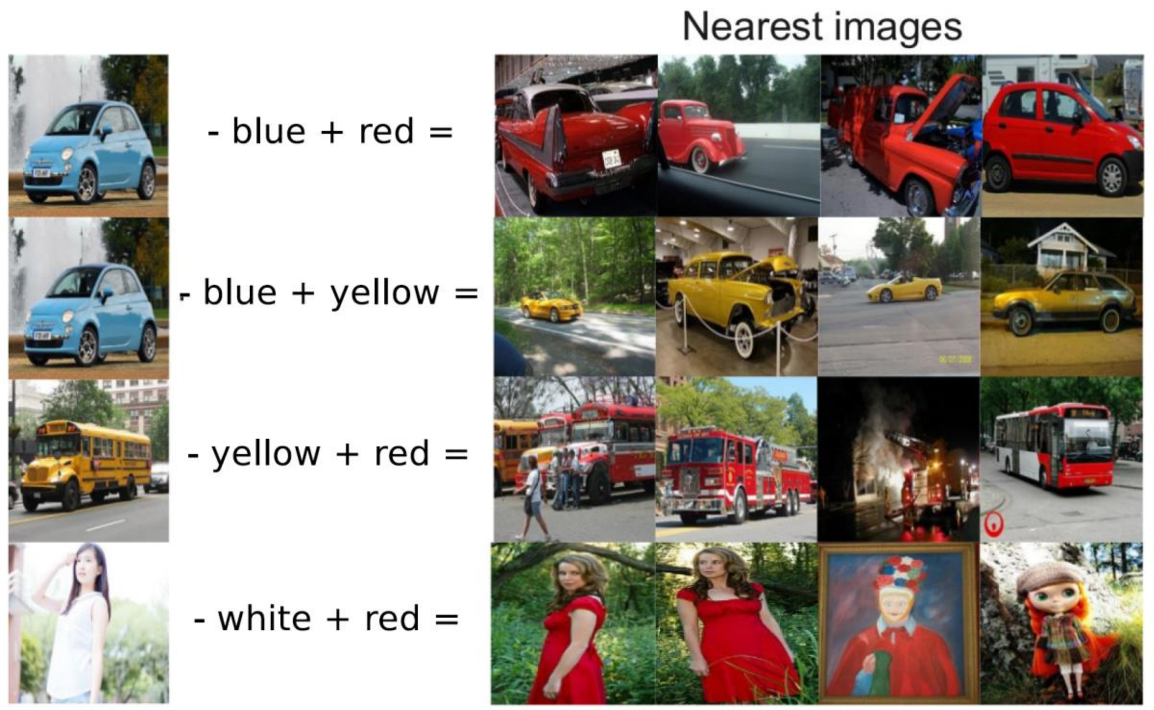
Audio-visual emotion recognition

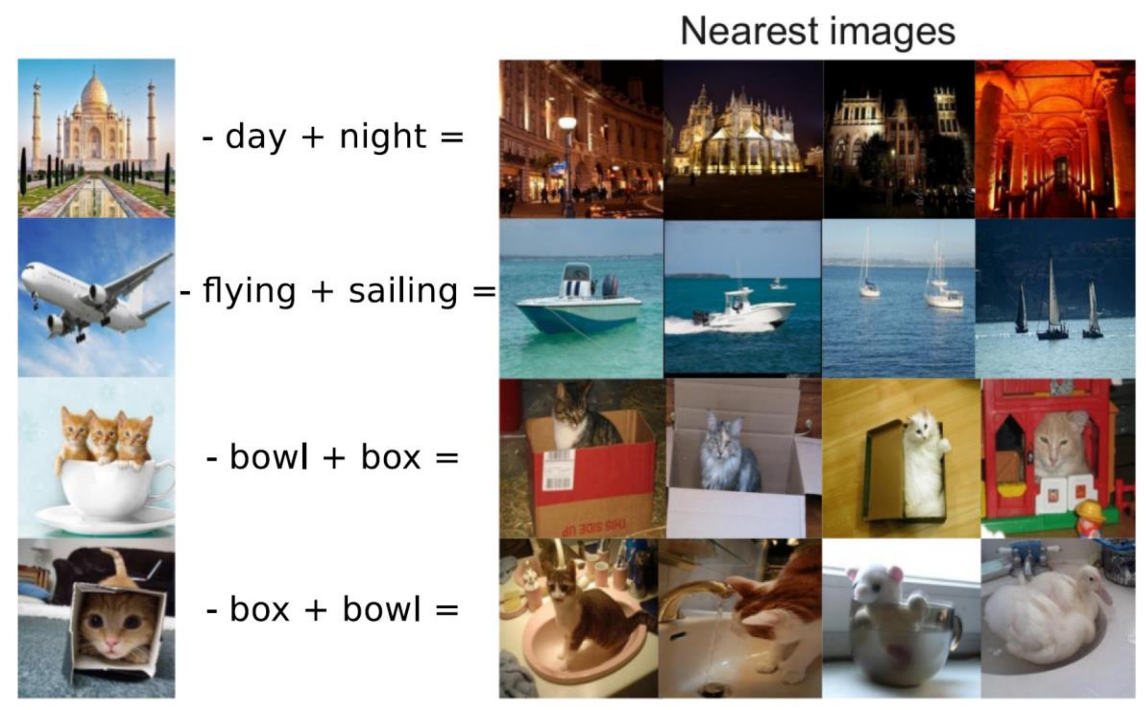
[Kim et al., ICASSP 2013]

- Deep Boltzmann Machine



Multimodal Vector Space Arithmetic



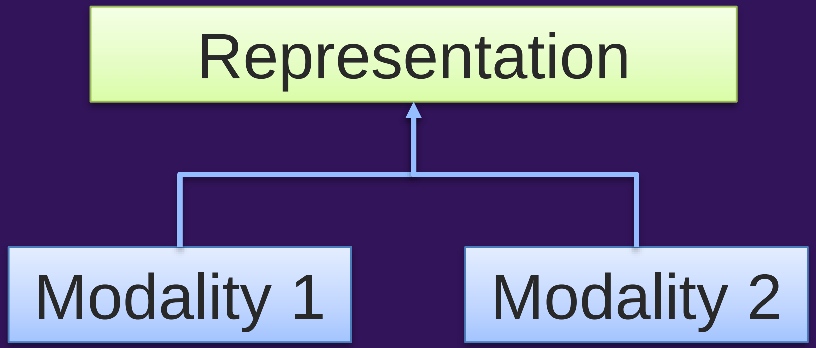


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

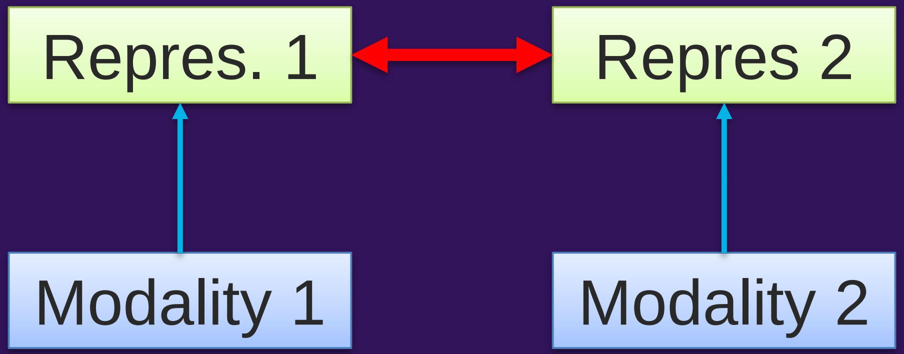
Core Challenge 1: Representation

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

A Joint representations:



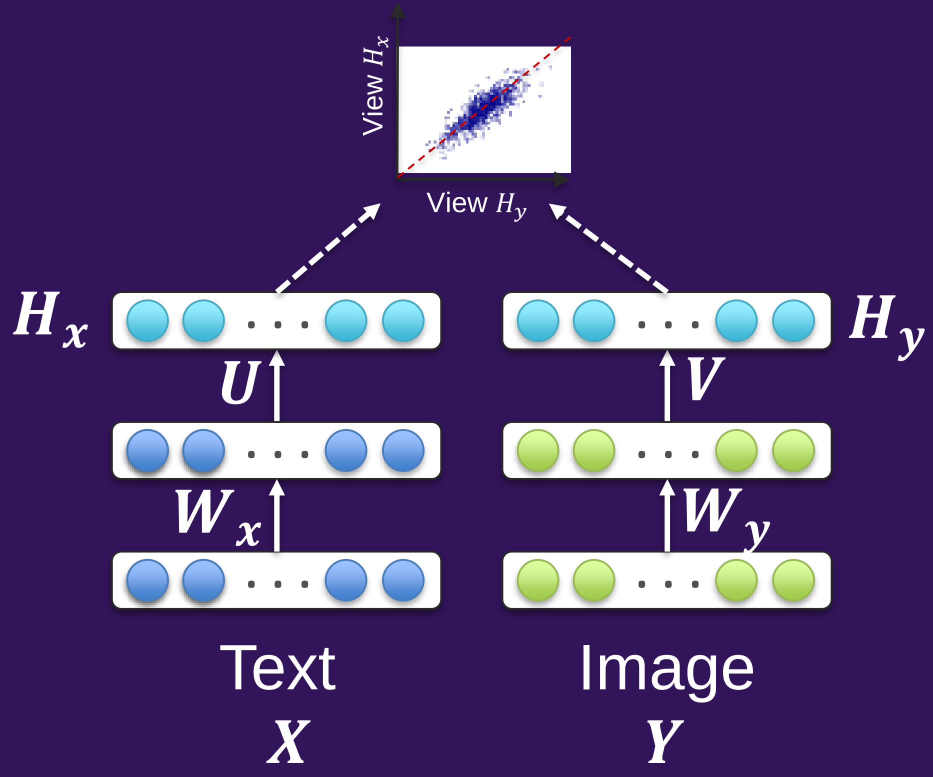
B Coordinated representations:



Coordinated Representation: Deep CCA

Learn linear projections that are maximally correlated:

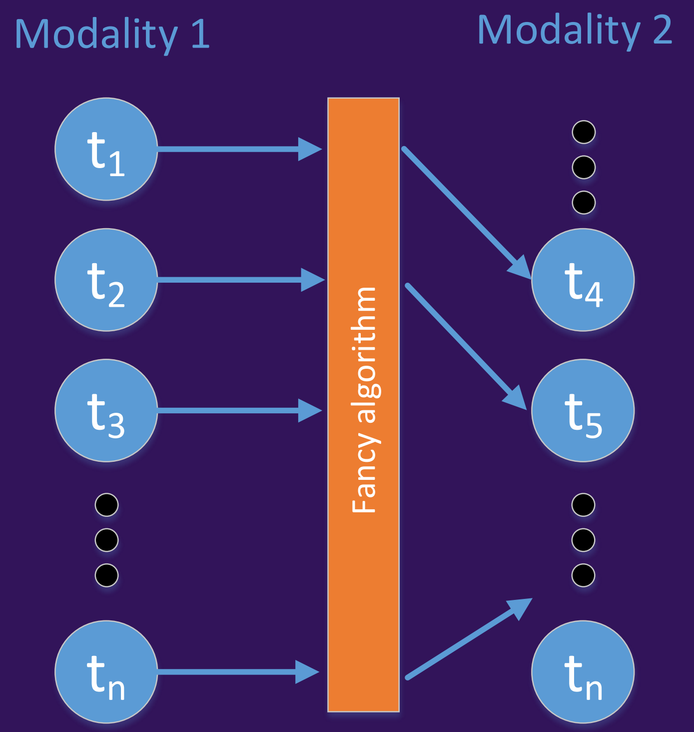
$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr\left(\boldsymbol{u}^T \boldsymbol{X}, \boldsymbol{v}^T \boldsymbol{Y}\right)$$



Andrew et al., ICML 2013

Core Challenge 2: Alignment

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



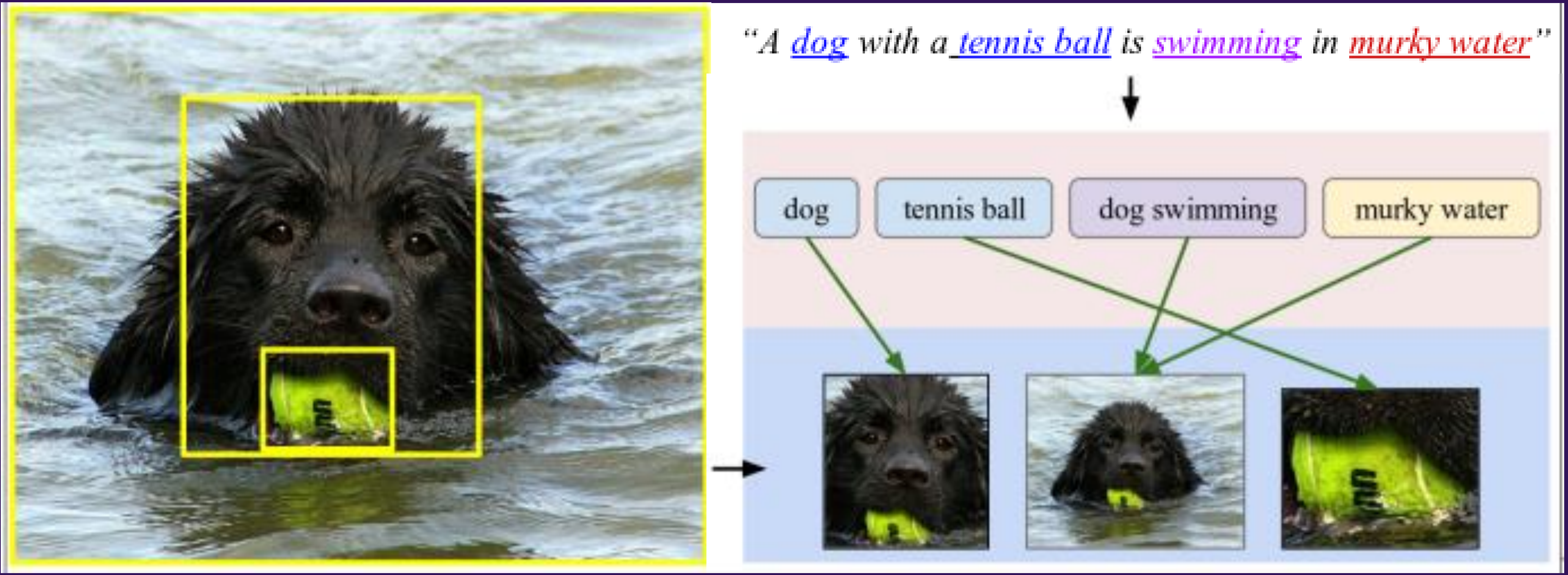
A **Explicit Alignment**

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

B **Implicit Alignment**

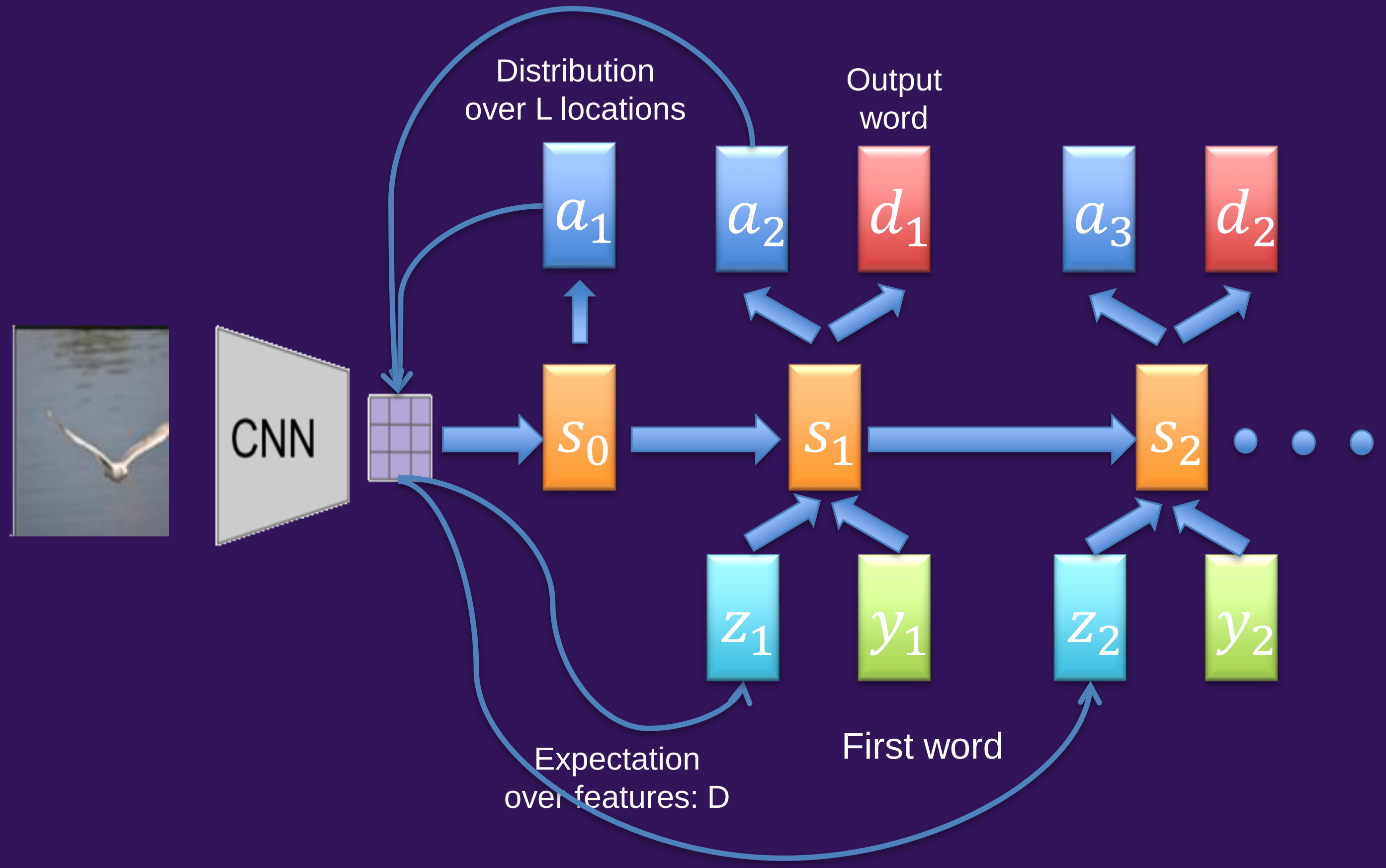
Uses internally latent alignment of modalities in order to better solve a different problem

Implicit Alignment



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

Attention Models for Image Captioning



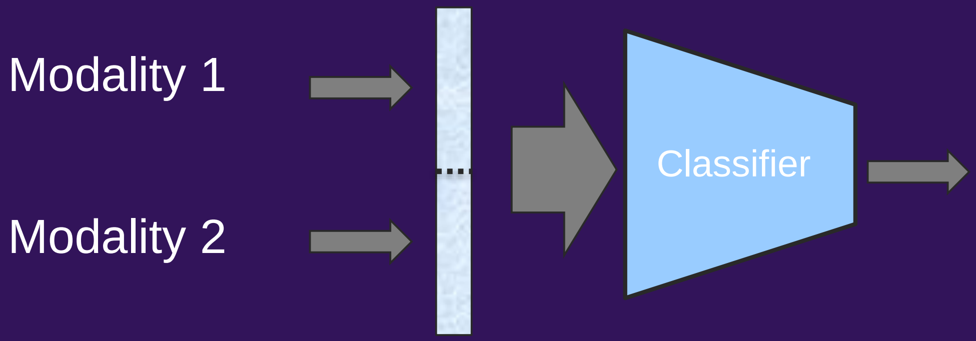
Xu et.al., ICML 2015

Core Challenge 3: Fusion

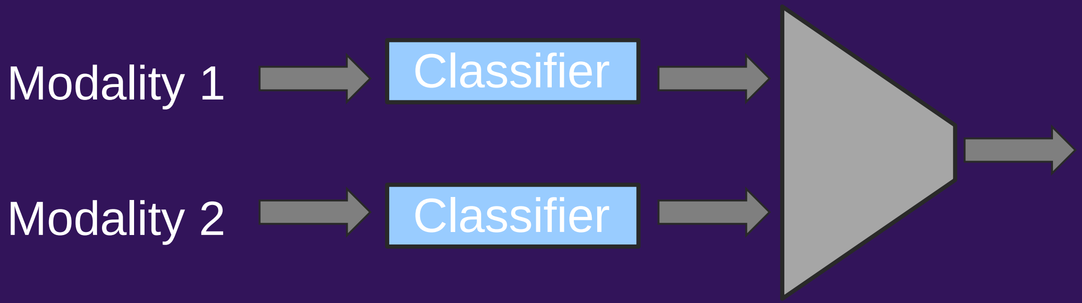
Definition: To join information from two or more modalities to perform a prediction task.

Ⓐ Model-Agnostic Approaches

1) Early Fusion



2) Late Fusion



Core Challenge 3: Fusion

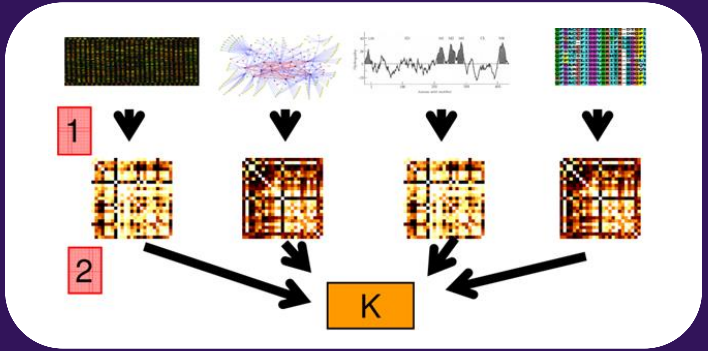
Definition: To join information from two or more modalities to perform a prediction task.

(B) Model-Based (Intermediate) Approaches

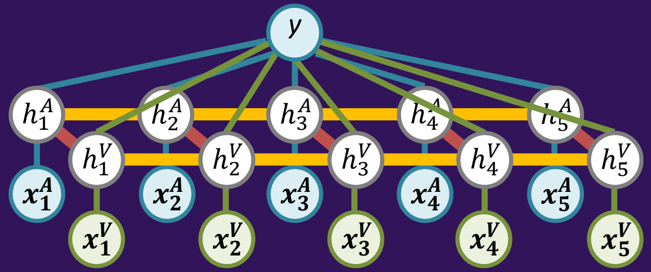
1) Deep neural networks

2) Kernel-based methods

3) Graphical models



Multiple kernel learning

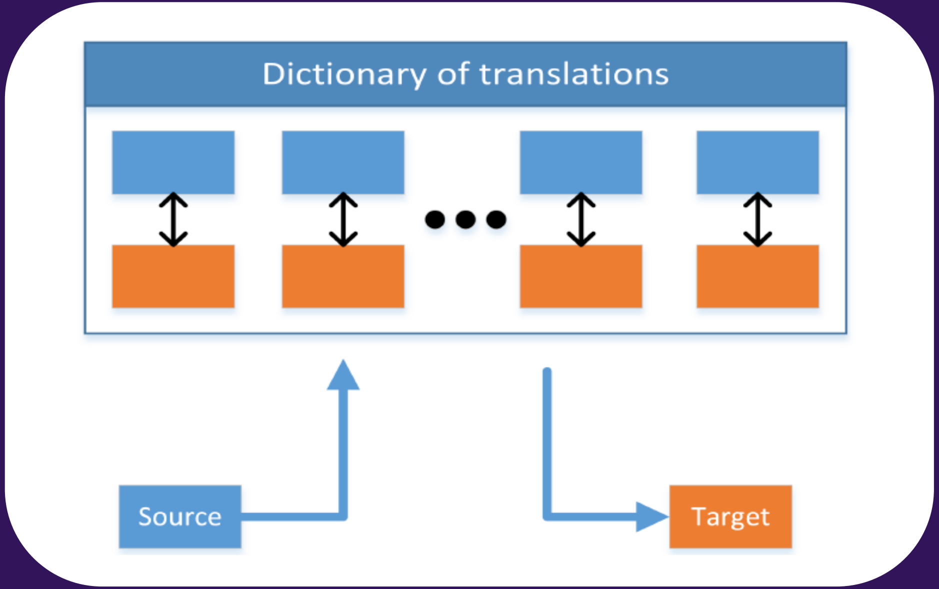


Multi-View Hidden CRF

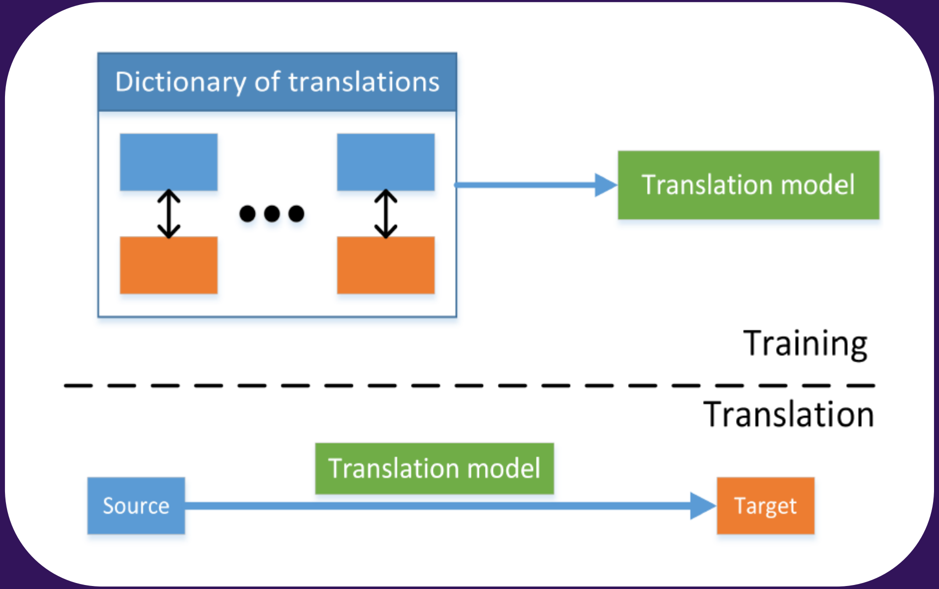
Core Challenge 4: Translation

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

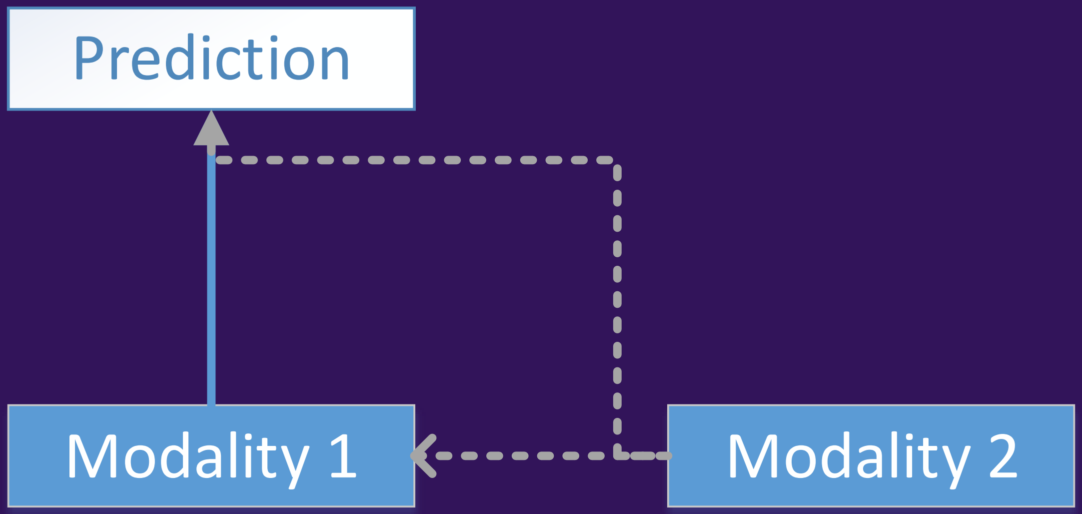
Visual gestures
(both speaker and listener gestures)

Transcriptions + Audio streams

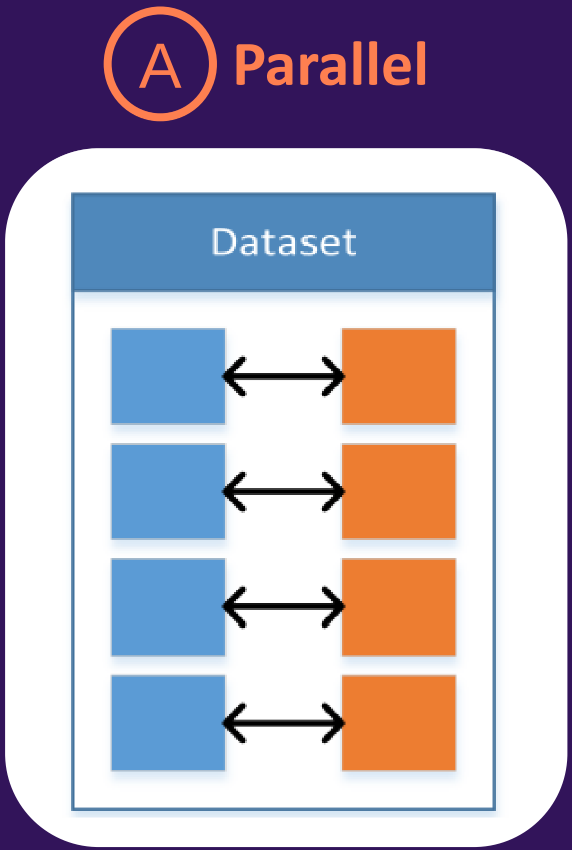
Marsella et al., Virtual character performance from speech, SIGGRAPH/Eurographics Symposium on Computer Animation, 2013

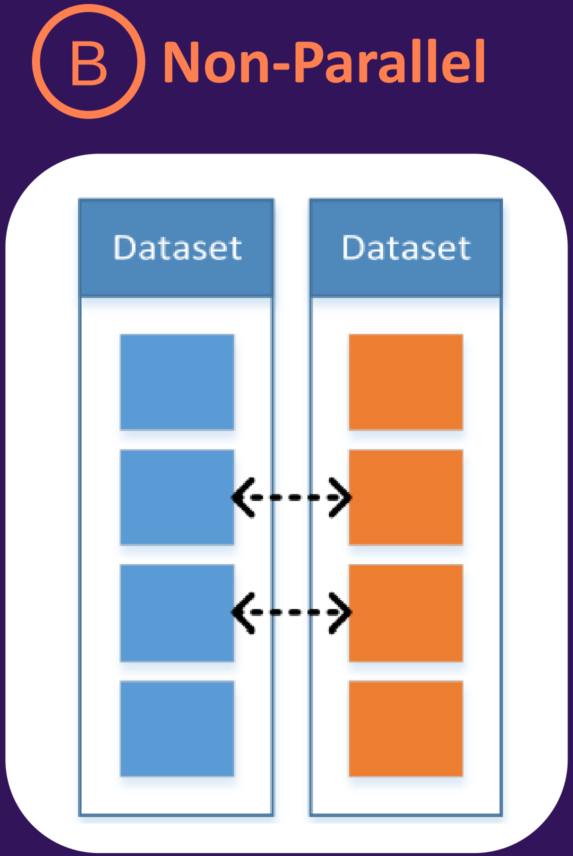
Core Challenge 5: Co-Learning

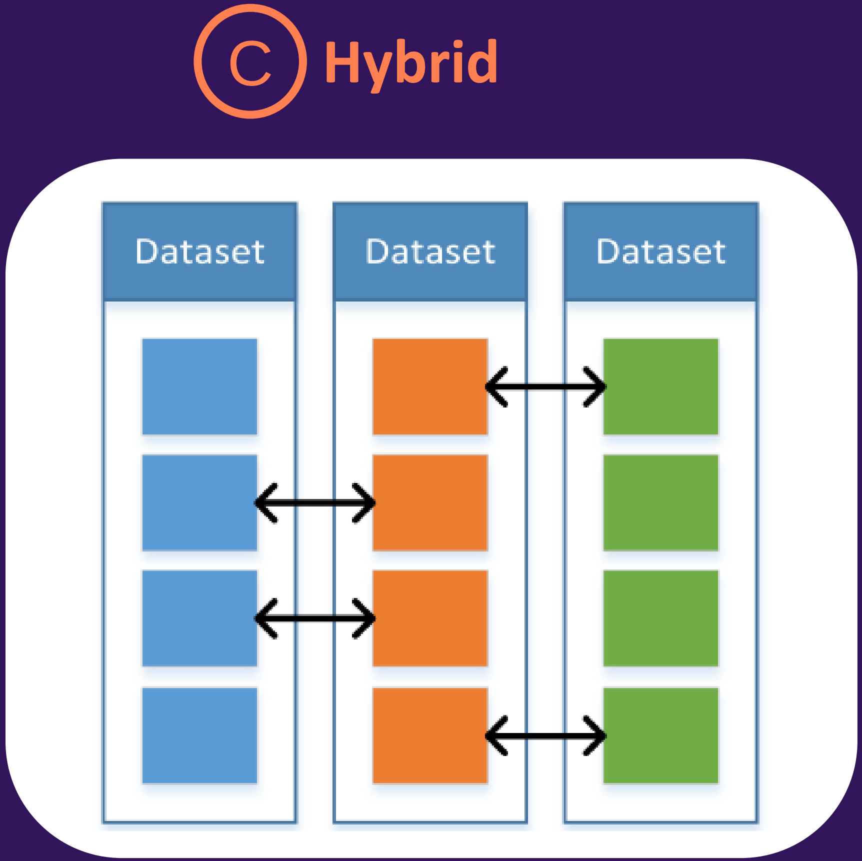
Definition: Transfer knowledge between modalities, including their representations and predictive models.



Core Challenge 5: Co-Learning







Taxonomy of Multimodal Research

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, https://arxiv.org/abs/1705.09406

Recent Progress in Multimodal ML

Representation → Multimodal Tensor Representation [ACL 2017, EMNLP 2017]

Alignment → Temporal Attention-Gated [CVPR 2017, ACM MM 2017]

Fusion → Multi-View Coupled LSTM [ECCV 2016]

Translation

Co-Learning

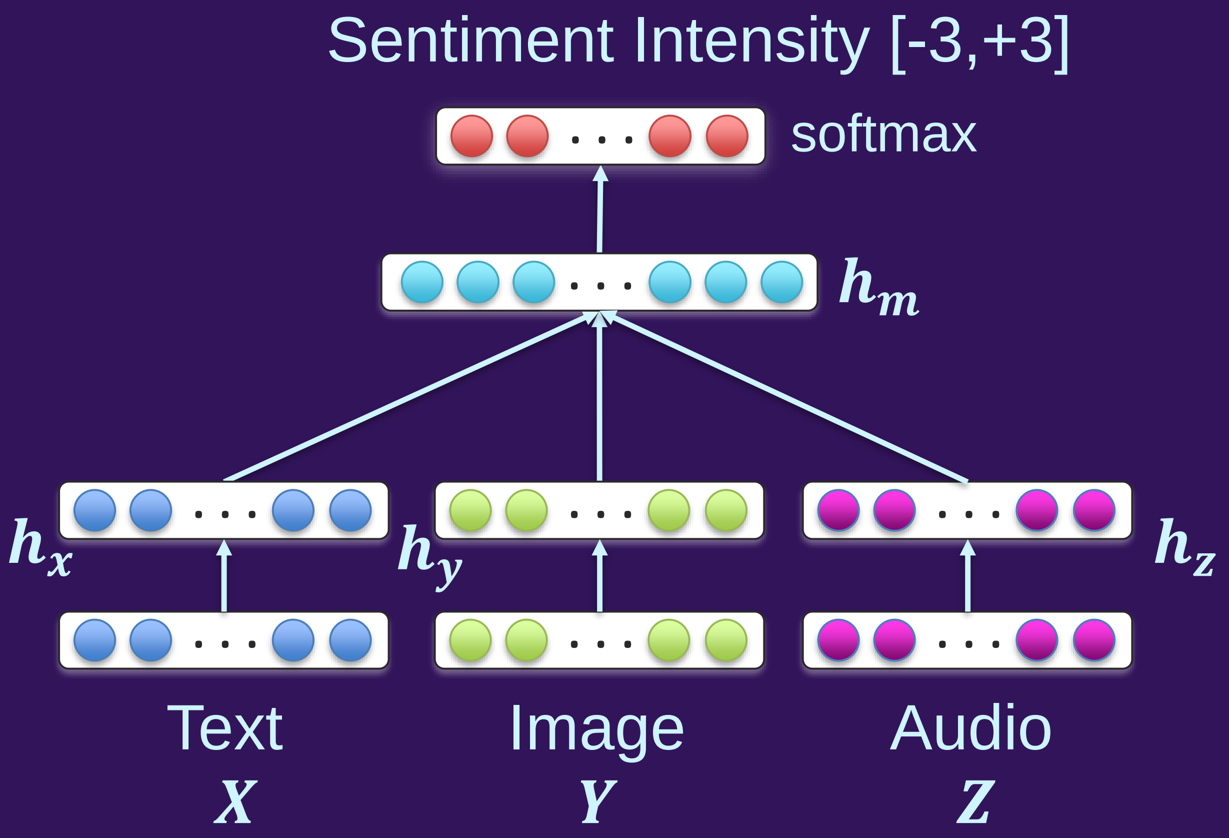
Multimodal Sentiment Analysis

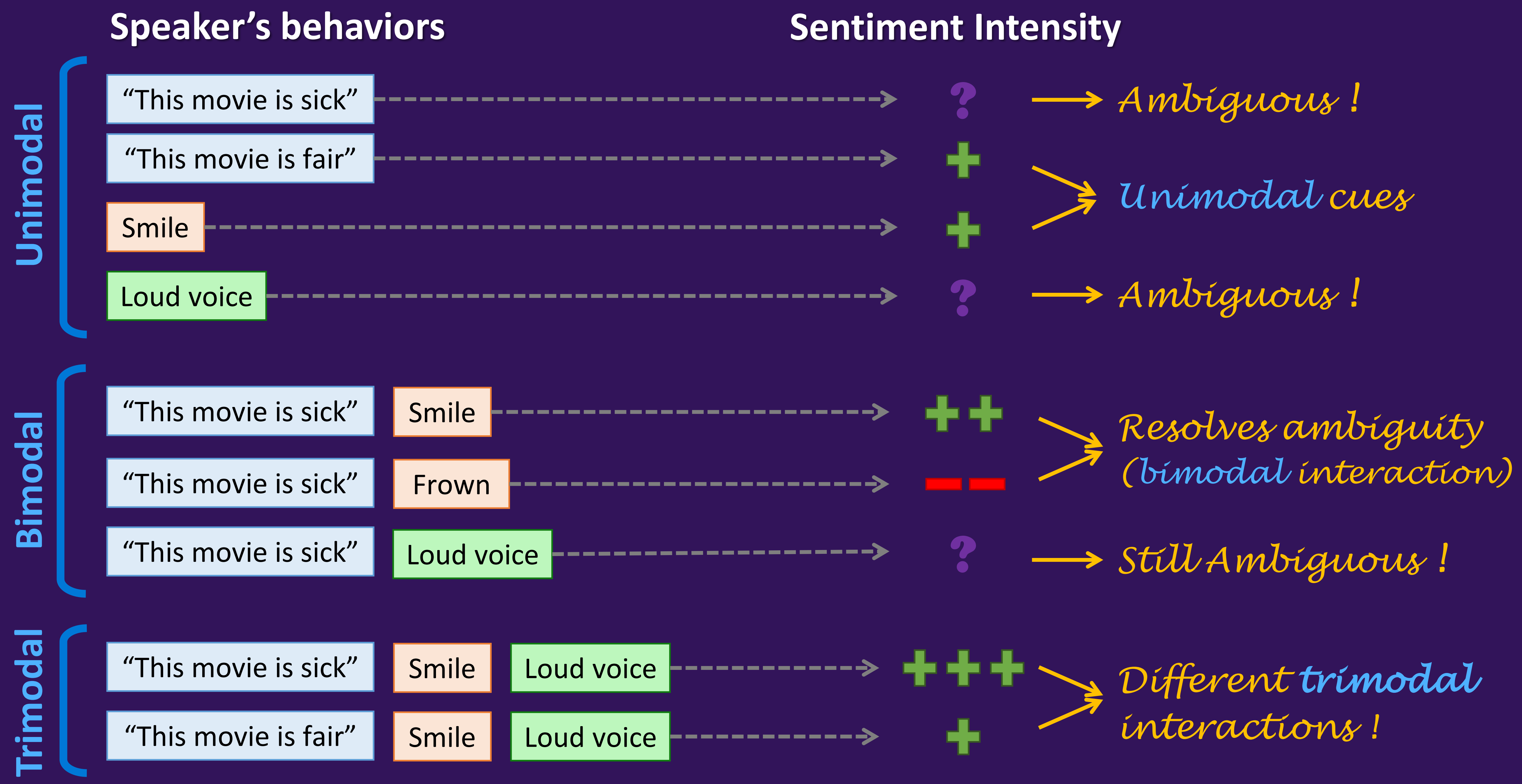
MOSI dataset (Zadeh et al, 2016)

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

Multimodal joint representation:

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



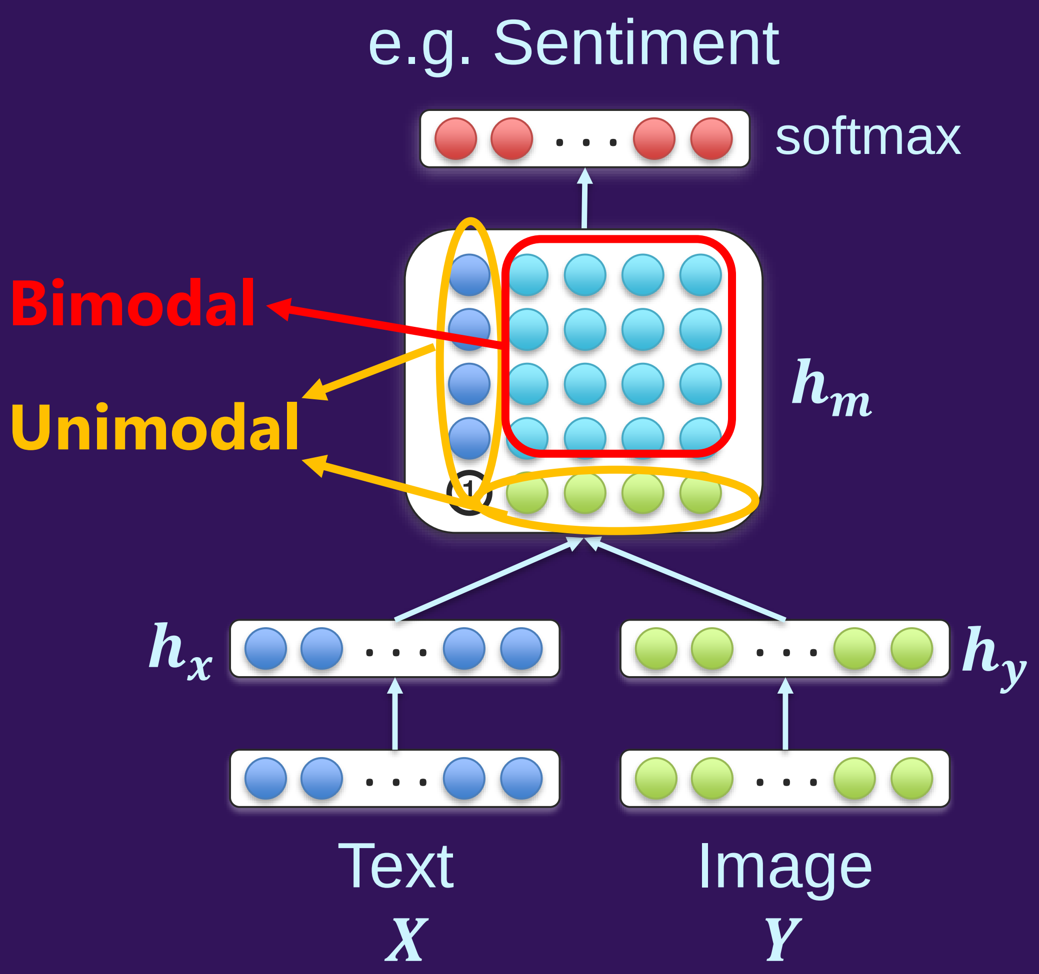

Speaker's behaviors
Sentiment Intensity
Unimodal
"This movie is sick"
?
Ambiguous !
"This movie is fair"
+
Unimodal cues
Smile
+
Loud voice
?
Ambiguous !
Bimodal
"This movie is sick"
Smile
++
Resolves ambiguity (bimodal interaction)
"This movie is sick"
Frown
− −
"This movie is sick"
Loud voice
?
Still Ambiguous !
Trimodal
"This movie is sick"
Smile
Loud voice
+++
Different trimodal interactions !
"This movie is fair"
Smile
Loud voice
+

Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

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

Important !



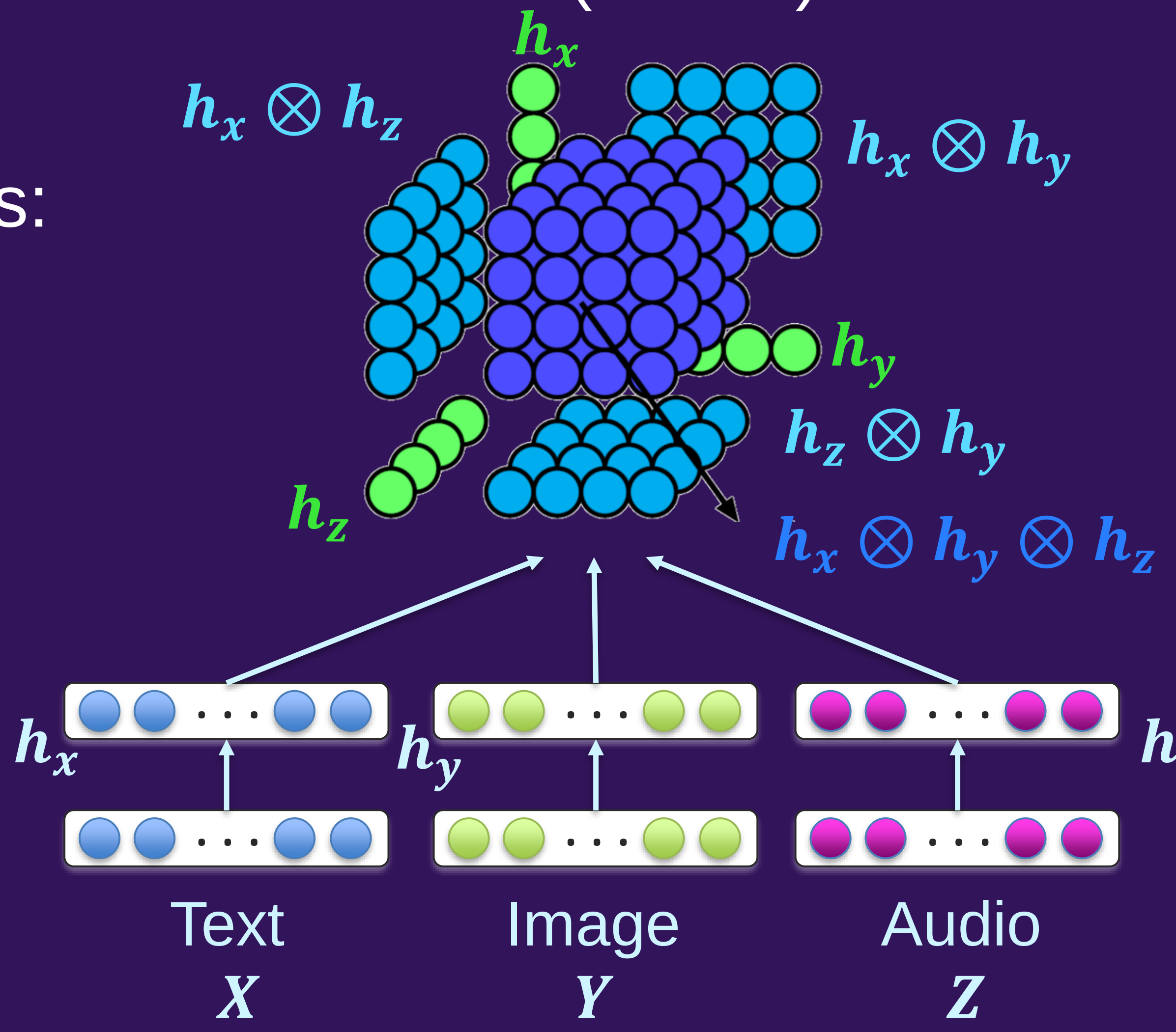
[Zadeh, Jones and Morency, **EMNLP 2017**]

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**]

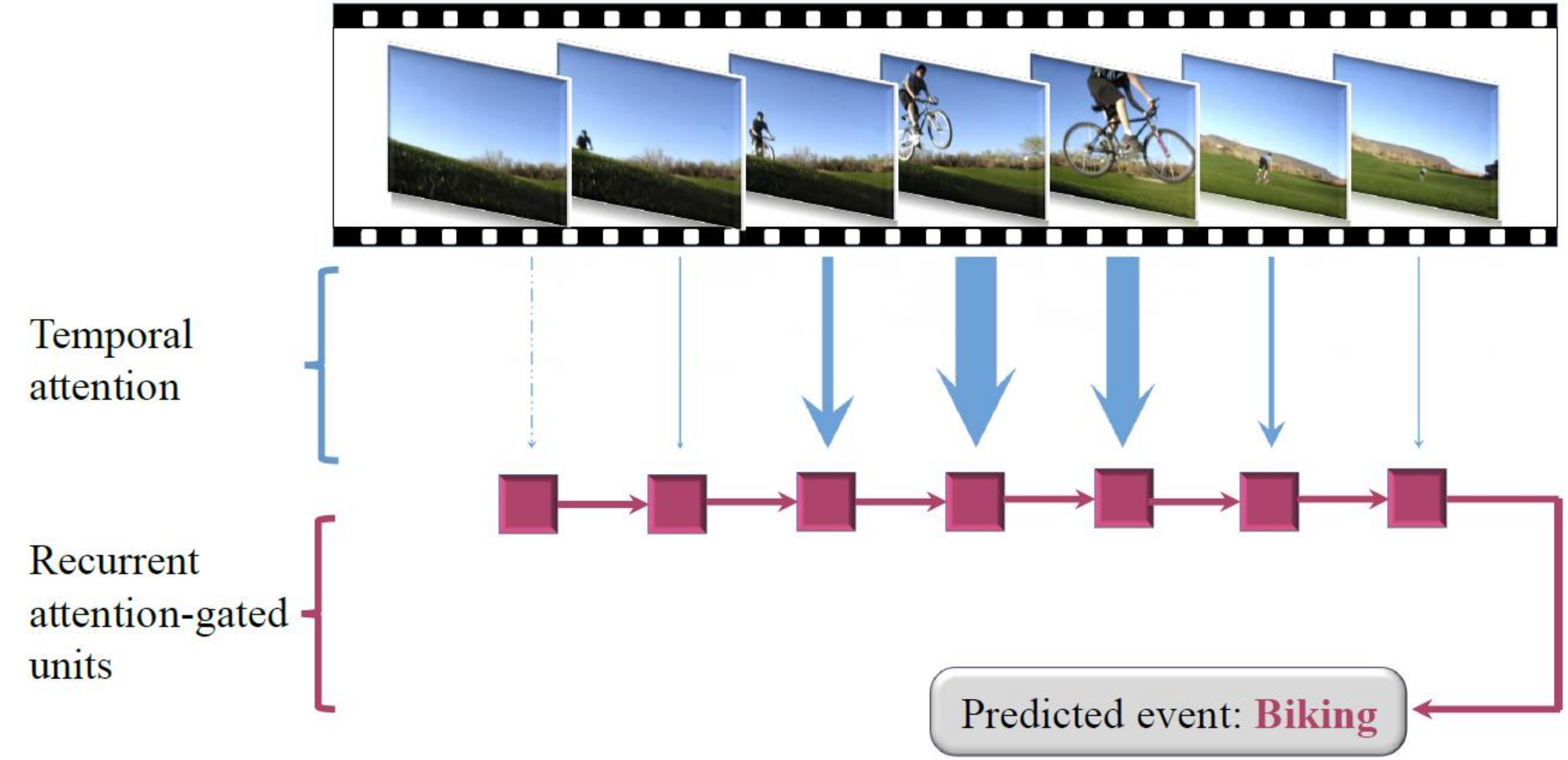
Experimental Results – MOSI Dataset

Multimodal Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	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}	$\uparrow$ 4.0	$\uparrow$ 2.7	$\uparrow$ 6.7	$\downarrow$ 0.23	$\uparrow$ 0.17

Improvement over State-Of-The-Art

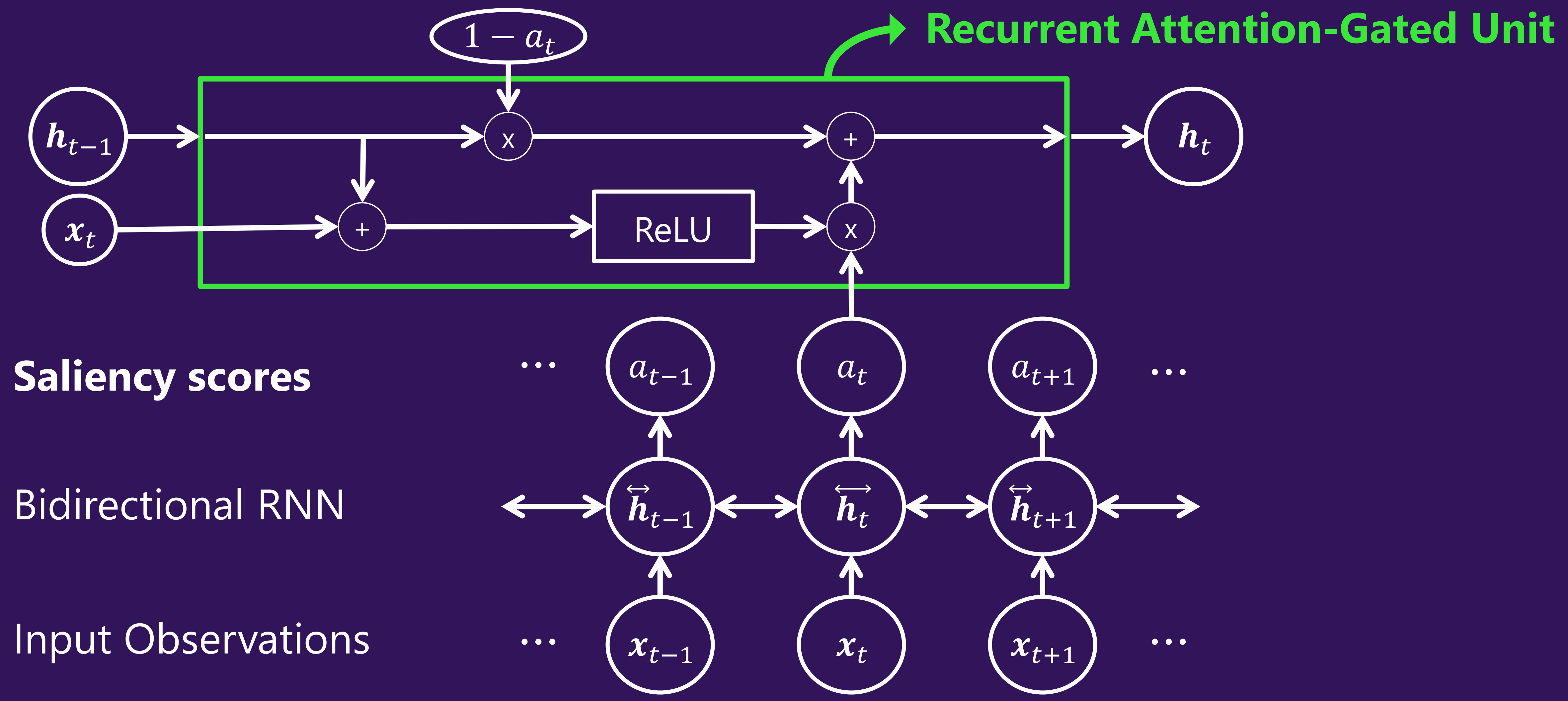
Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
$\mathrm{TFN}_{language}$	74.8	75.6	38.5	0.99	0.61
TFN_{visual}	66.8	70.4	30.4	1.13	0.48
$\mathrm{TFN}_{acoustic}$	65.1	67.3	27.5	1.23	0.36
$\mathrm{TFN}_{bimodal}$	75.2	76.0	39.6	0.92	0.65
$\mathrm{TFN}_{trimodal}$	74.5	75.0	38.9	0.93	0.65
$\mathrm{TFN}_{notrimodal}$	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

Temporal Attention in Videos

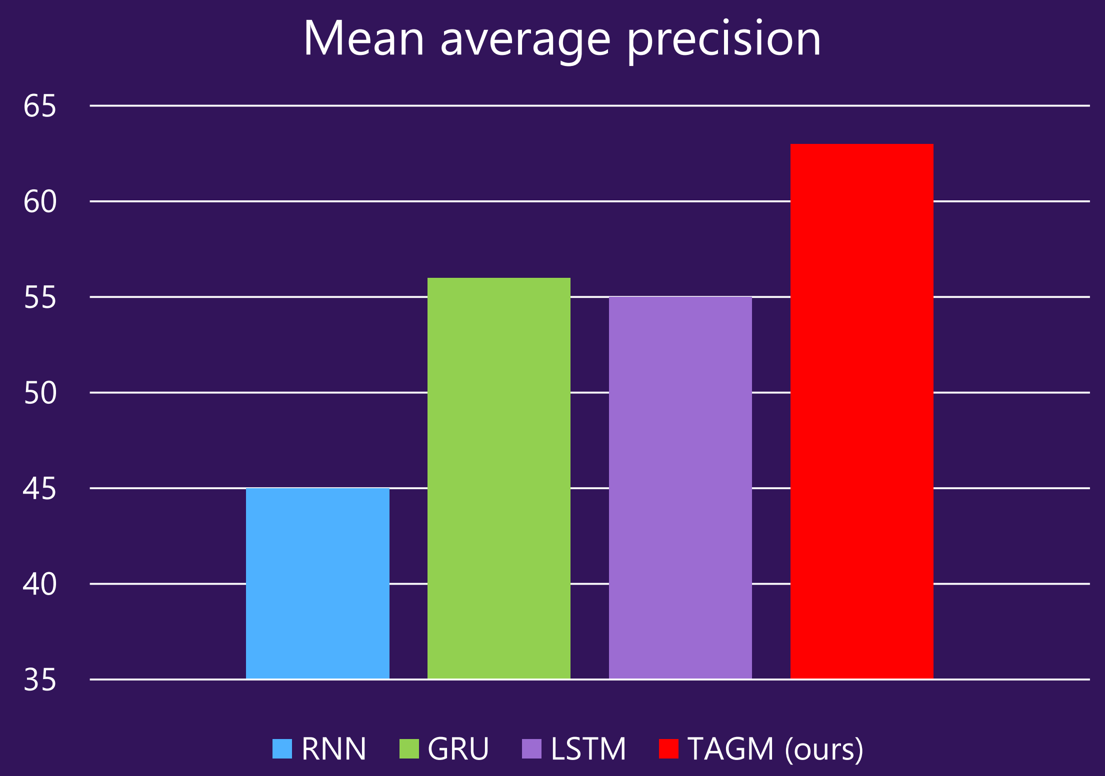


Pei, Baltrušaitis, Tax and Morency. Temporal Attention-Gated Model for Robust Sequence Classification, **CVPR**, *2017*

Temporal Attention-Gated Model (TAGM)

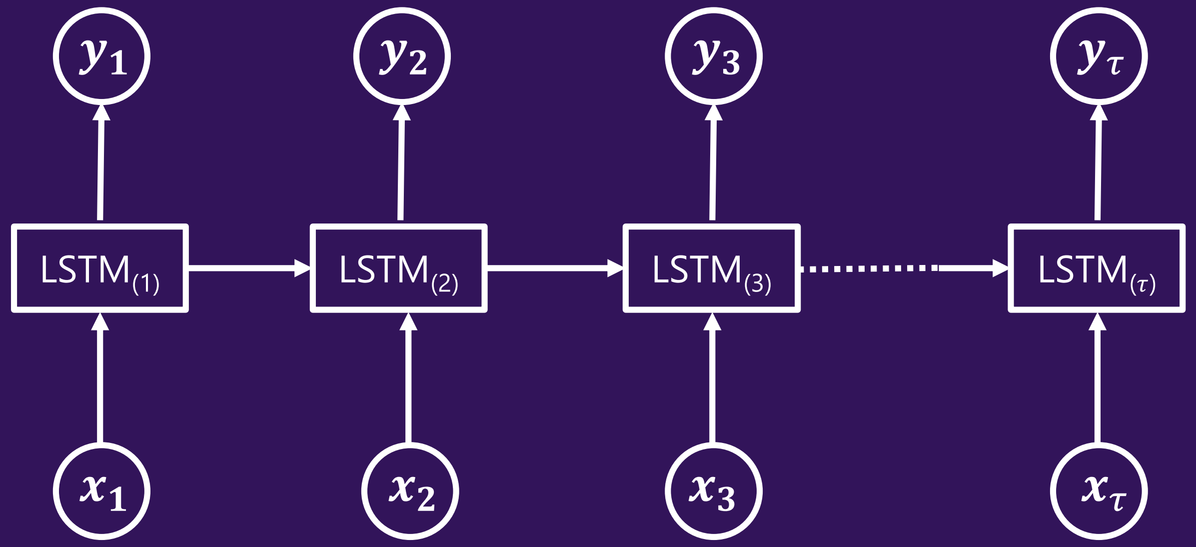


Experimental Results – CCV Dataset

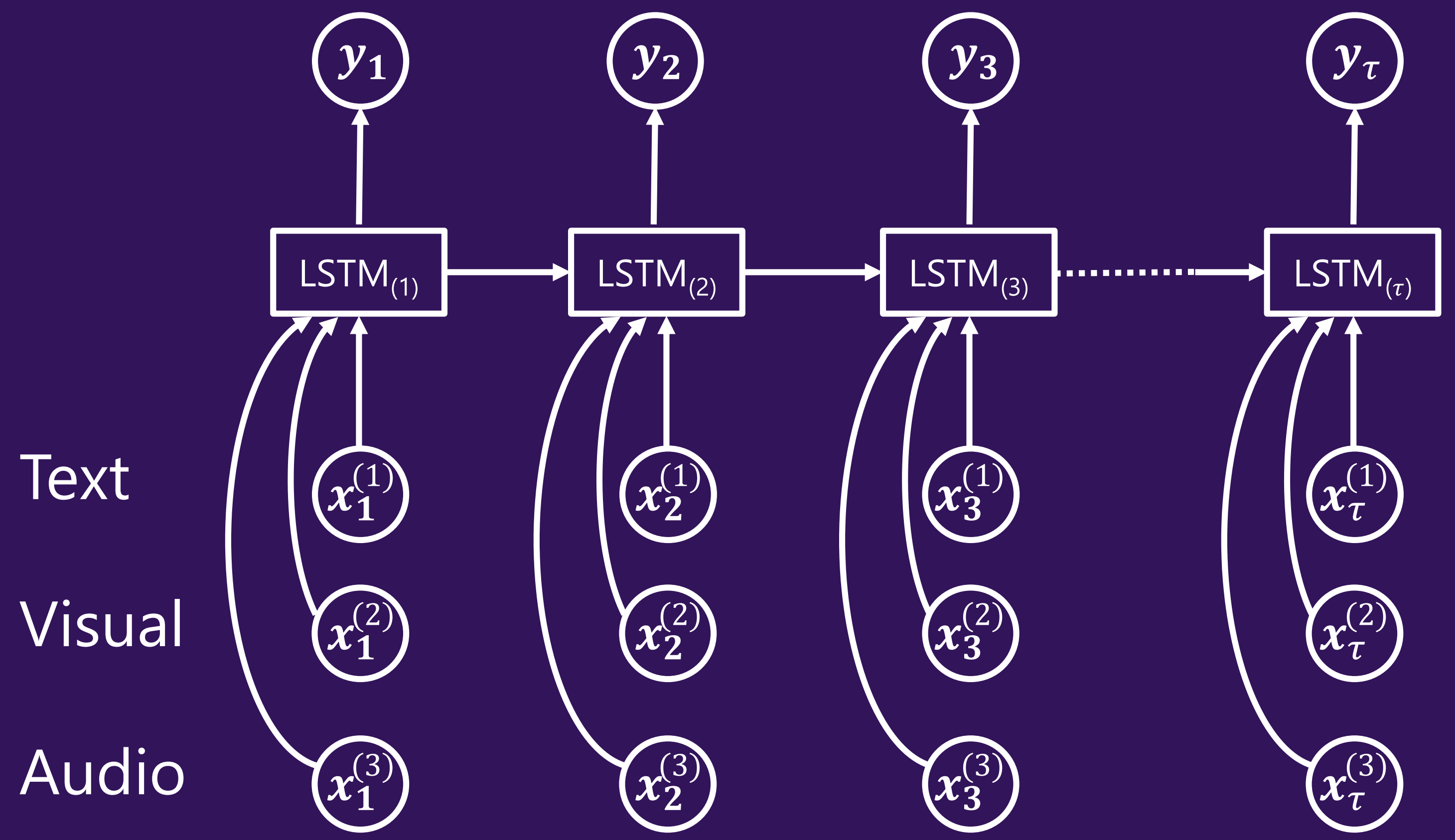


Pei, Baltrušaitis, Tax and Morency. Temporal Attention-Gated Model for Robust Sequence Classification, **CVPR**, *2017*

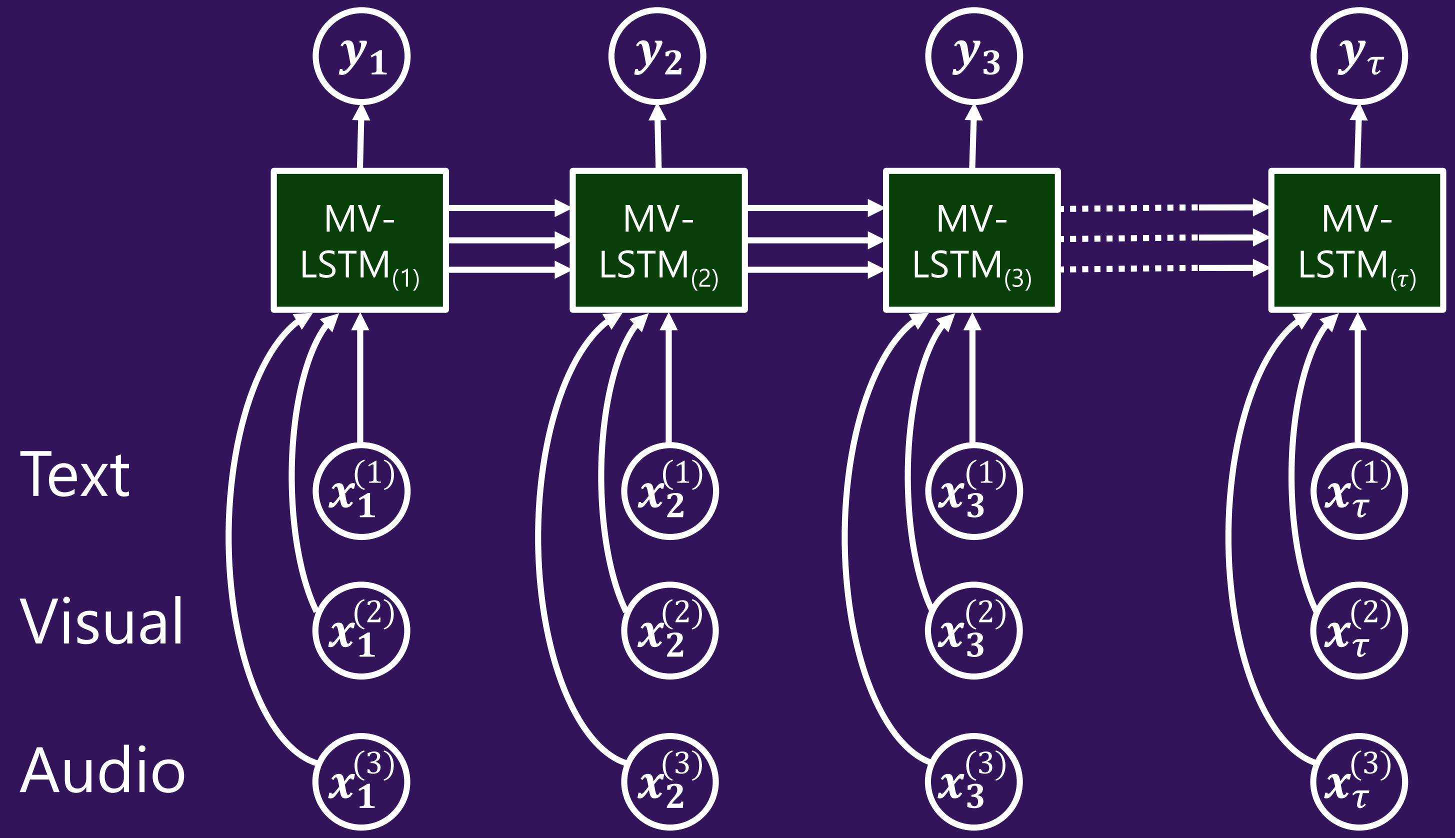
Sequence Modeling with LSTM



Multimodal Sequence Modeling – Early Fusion

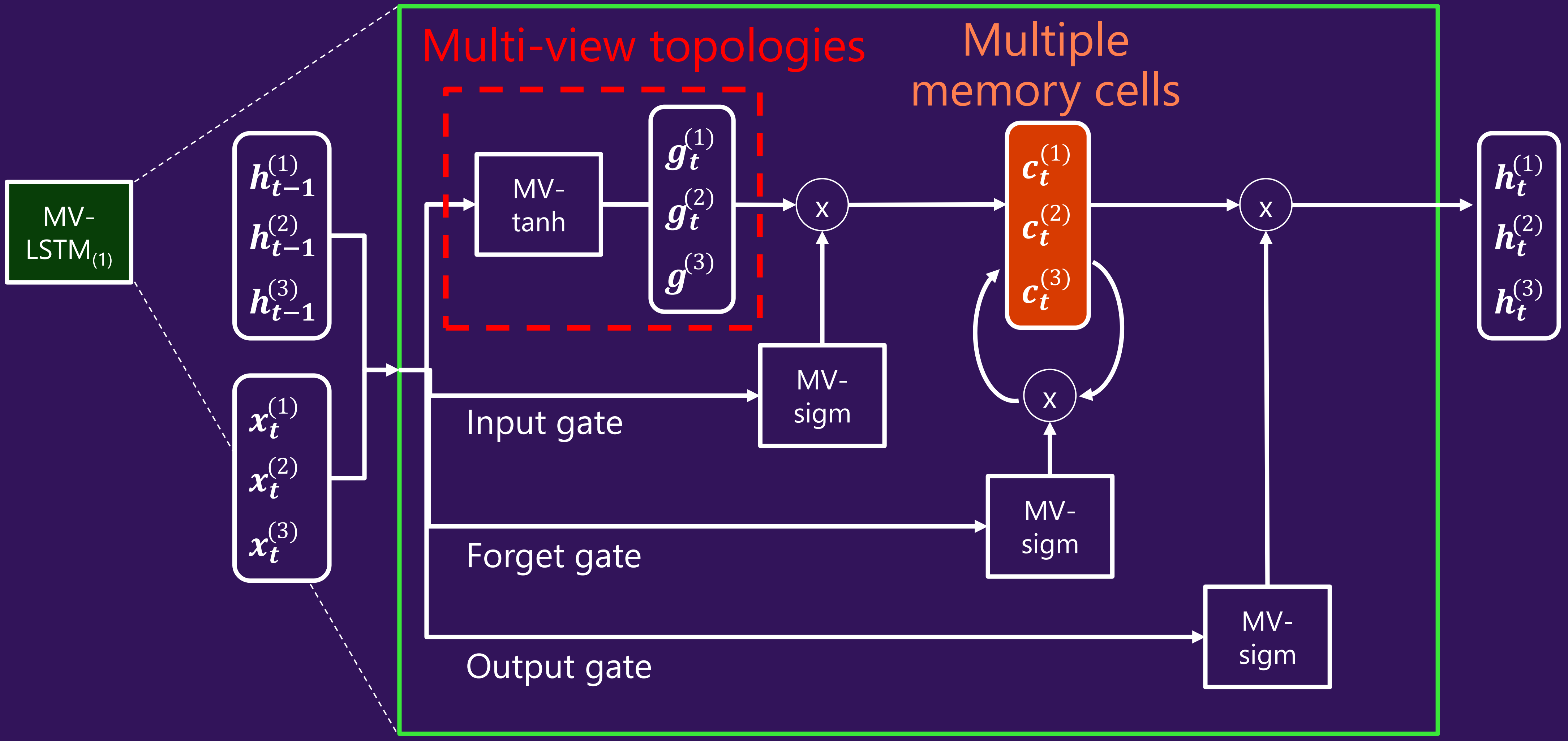


Multi-View Long Short-Term Memory

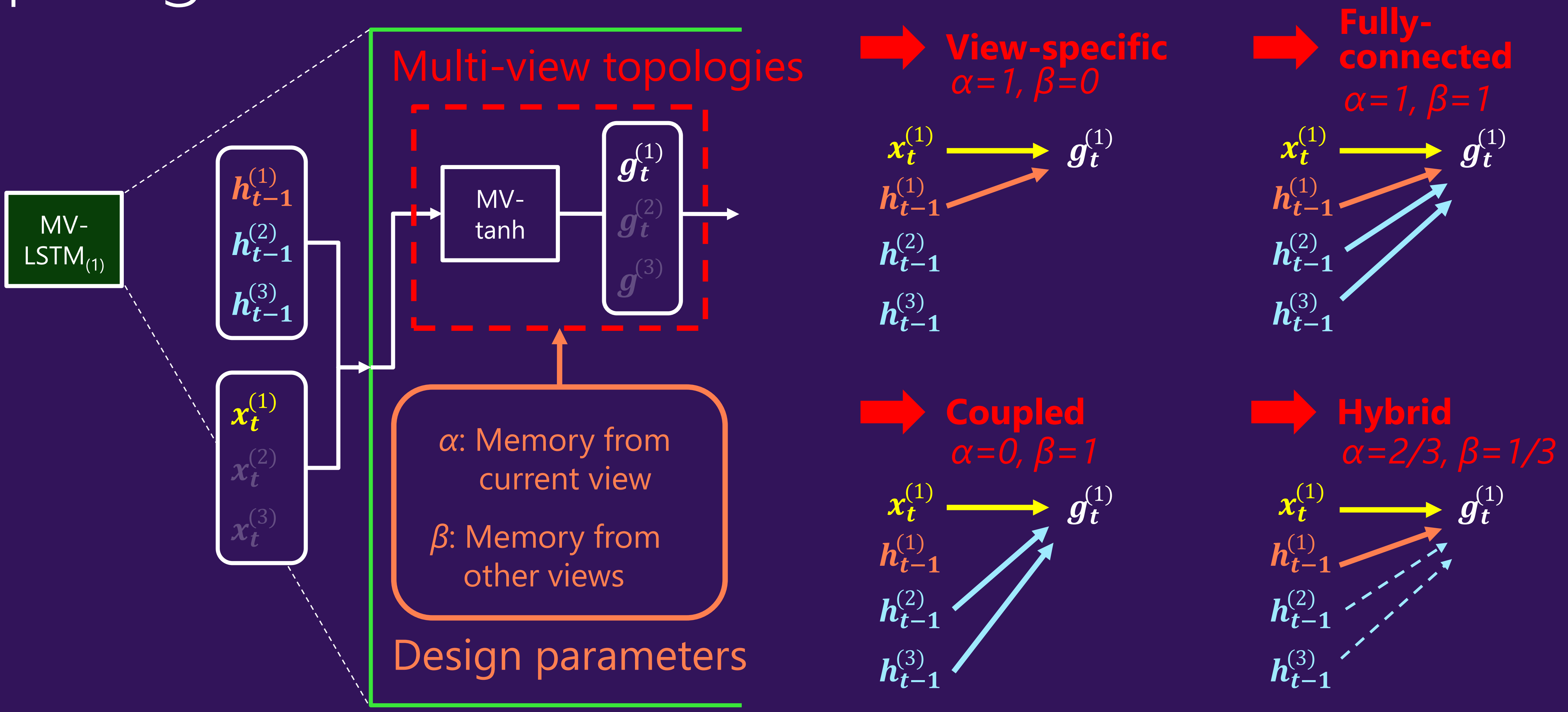


[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

Multi-View Long Short-Term Memory



Topologies for Multi-View LSTM



Experimental Results

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]

Multimodal Machine Learning

Representation

Alignment

Fusion

Translation

Co-Learning

Multimodal Machine Learning: A Survey and Taxonomy

https://arxiv.org/abs/1705.09406

→ Multimodal Tensor Representation
[ACL 2017, EMNLP 2017]

→ Temporal Attention-Gated
[CVPR 2017, ACM MM 2017]

→ Multi-View LSTM
[ECCV 2016]