Language Technologies Institute

Carnegie Mellon University

Multimodal Affective Computing

Lecture 13: Multimodal Deep Learning

Louis-Philippe Morency
Jeffrey Girard

Originally developed with help from Stefan Scherer and Tadas Baltrušaitis

Outline

- Multimodal core challenges - review
- Multimodal representations
  - Joint and coordinated representations
    - Multimodal autoencoder & tensor fusion
    - Deep canonical correlation analysis
- Multimodal alignment
  - Implicit and explicit alignment
    - Dynamic time warping
    - Attention models
- Multimodal fusion
  - Multi-view recurrent network
  - Memory fusion networks

Upcoming Lectures

Classes	Tuesday	Thursday
Week 13 **4/09 & 4/11**	Multimodal deep learning • Multimodal representations • Attention and modality alignment • Temporal and multimodal fusion	NO CLASS
Week 14 **4/16 & 4/18**	Multimodal Behavior Generation • Guest lecture: Prof. Nakano • Generation based on user's attitude • Robot and virtual humans	Discussion (generation) • Jiang Liu • Ankit Shah
Week 15 **4/23 & 4/25**	Multimodal applications • Assessment in the clinical process • Biomarkers and behavioral indicators • Validation in the medical sciences	Discussion (applications) • Mingtong Zhang • Mahmoud Al Ismail
Week 16 **4/30 & 5/02** ***final report***	NO CLASS	Final presentations

Multimodal Machine Learning: Core Technical Challenges

Core Challenges in "Deep" Multimodal ML

Representation

Alignment

Fusion

Translation

Co-Learning

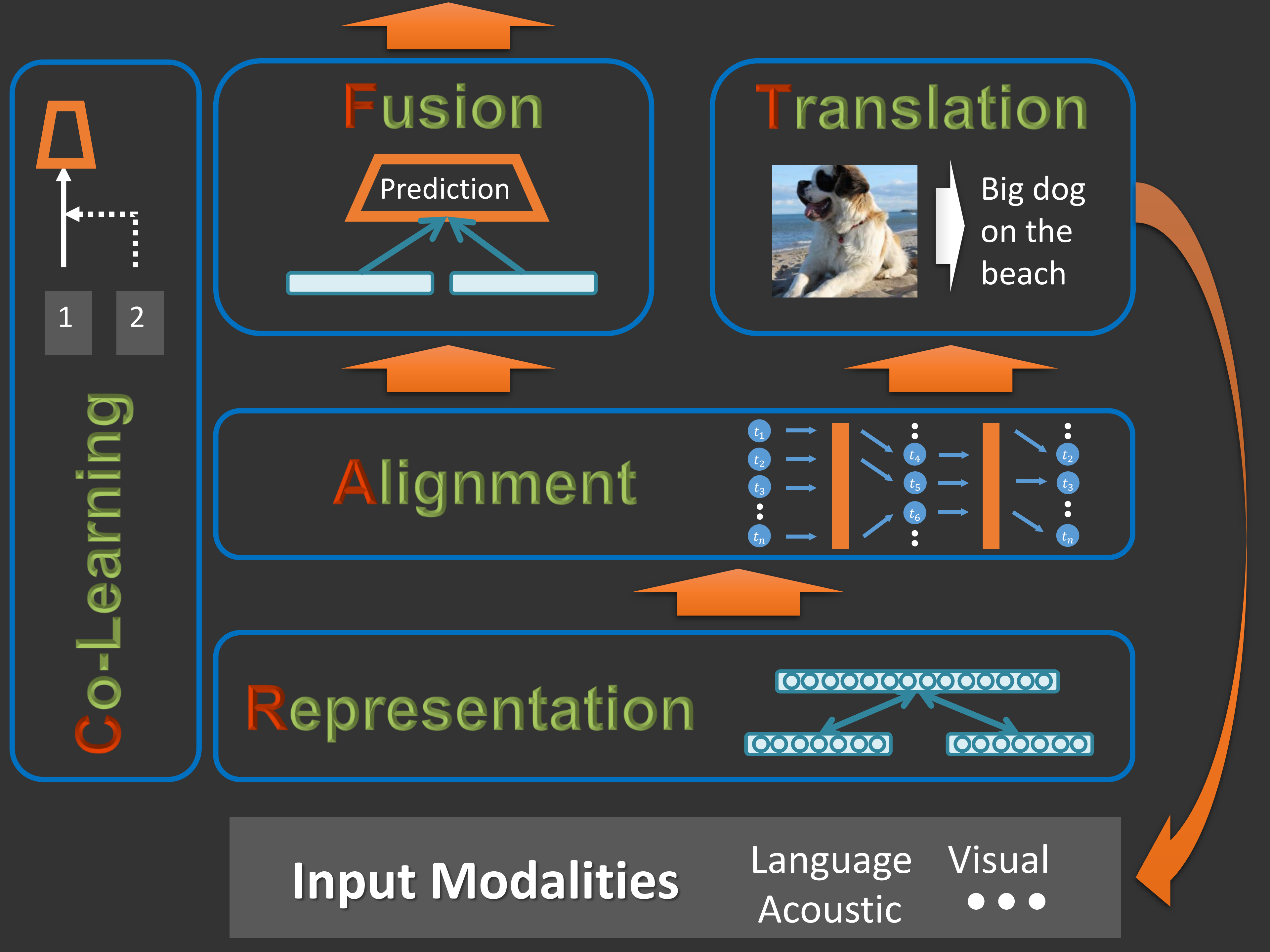
Multimodal Machine Learning: A Survey and Taxonomy

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.


Fusion
Prediction
Translation
Big dog on the beach
Alignment
t_1 t_2 t_3 t_n
t_4 t_5 t_6
t_2 t_3 t_n
Representation
Co-Learning
1
2
Input Modalities
Language
Acoustic
Visual
• • •

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

Real world tasks tackled by MMML

- 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

"man in black shirt is playing guitar." "construction worker in orange safety vest is working on road." "two young girls are playing with lego toy." "boy is doing backflip on wakeboard."

(a) answer-phone (a) get-out-car (a) fight-person (b) push-up (b) cartwheel

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	✓				✓

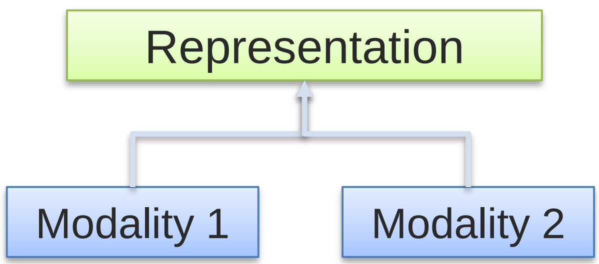
Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

Multimodal Representations

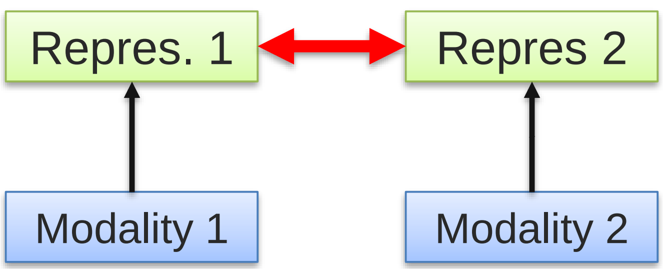
Core Challenge: Representation

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

Ⓐ Joint representations:

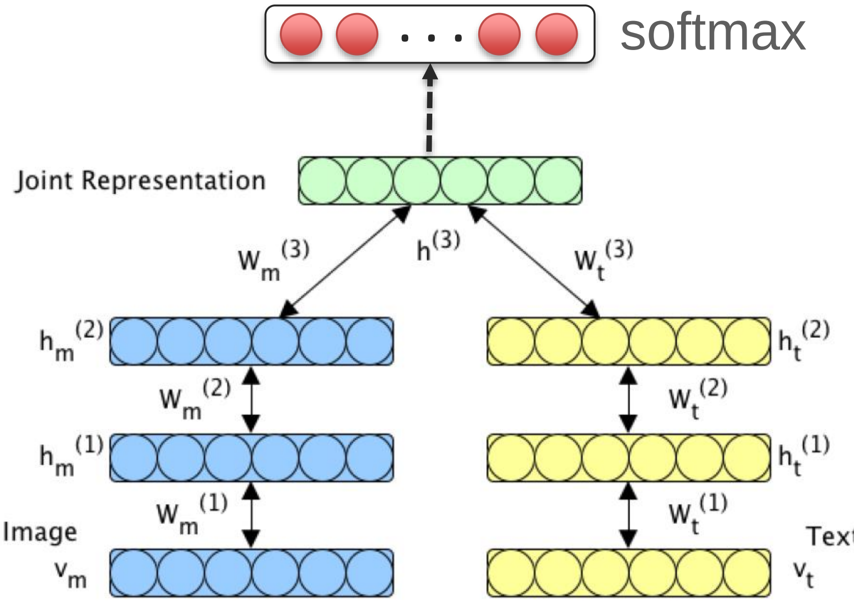


Ⓑ Coordinated representations:



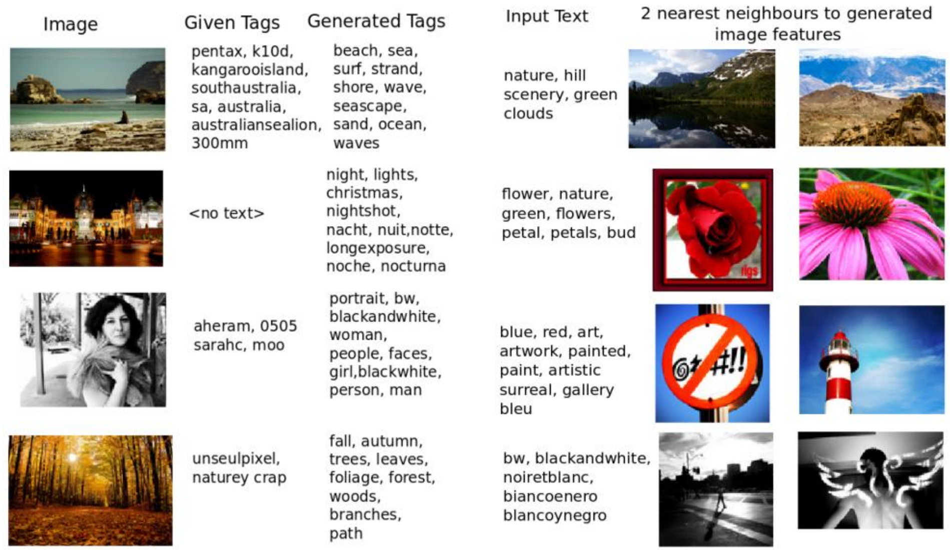
Deep Multimodal Boltzmann machines

- Generative model
- Individual modalities trained like a DBN
- Multimodal representation trained using Variational approaches
- Used for image tagging and cross-media retrieval
- Reconstruction of one modality from another is a bit more “natural” than in autoencoder representation
- Can actually sample text and images



[Srivastava and Salakhutdinov, Multimodal Learning with Deep Boltzmann Machines, 2012, 2014]

Deep Multimodal Boltzmann machines

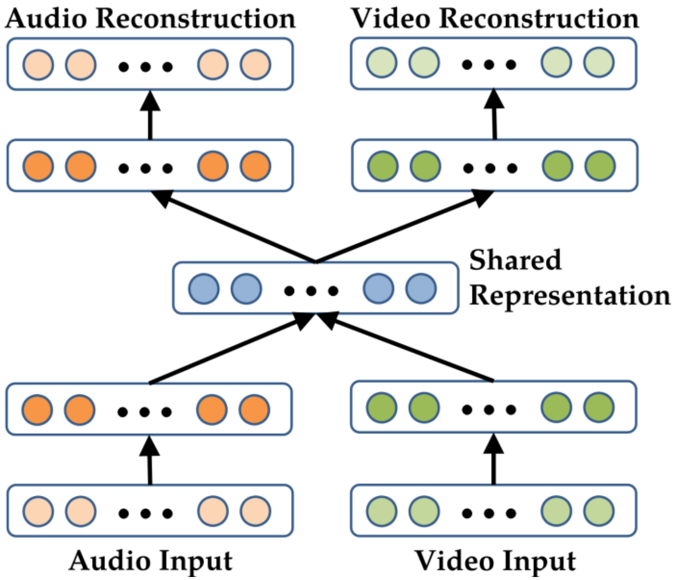


Model	MAP	Prec@50
Random	0.124	0.124
SVM (Huiskes et al., 2010)	0.475	0.758
LDA (Huiskes et al., 2010)	0.492	0.754
DBM	0.526 ± 0.007	0.791 ± 0.008
DBM (using unlabelled data)	**0.585** ± 0.004	**0.836** ± 0.004

Srivastava and Salakhutdinov, “Multimodal Learning with Deep Boltzmann Machines”, NIPS 2012

Deep Multimodal autoencoders

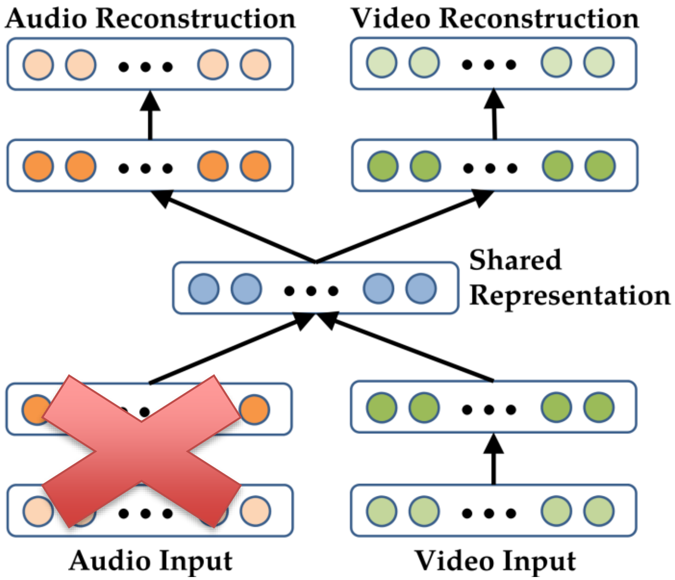
- A deep representation learning approach
- A bimodal auto-encoder
  - Used for Audio-visual speech recognition



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

Deep Multimodal autoencoders - training

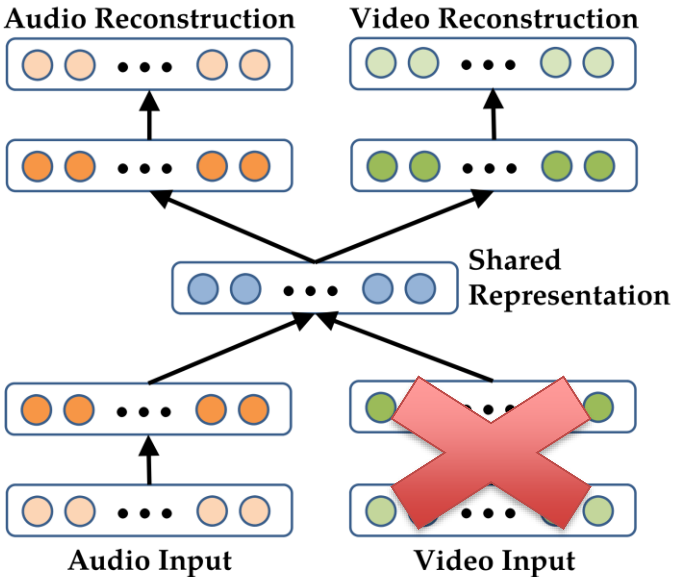
- Individual modalities can be pretrained
  - RBMs
  - Denoising Autoencoders
- To train the model to reconstruct the other modality
  - Use both
  - Remove audio



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

Deep Multimodal autoencoders - training

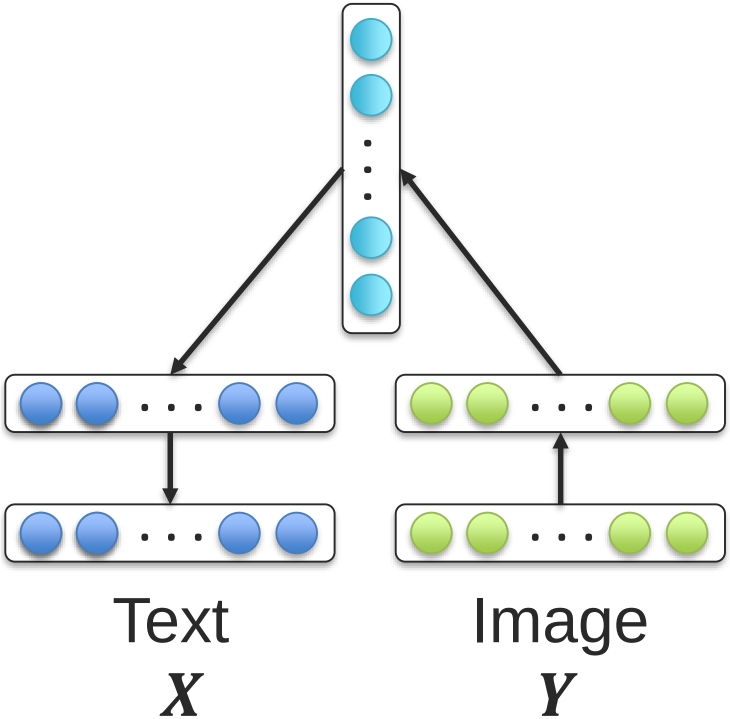
- Individual modalities can be pretrained
  - RBMs
  - Denoising Autoencoders
- To train the model to reconstruct the other modality
  - Use both
  - Remove audio
  - Remove video



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

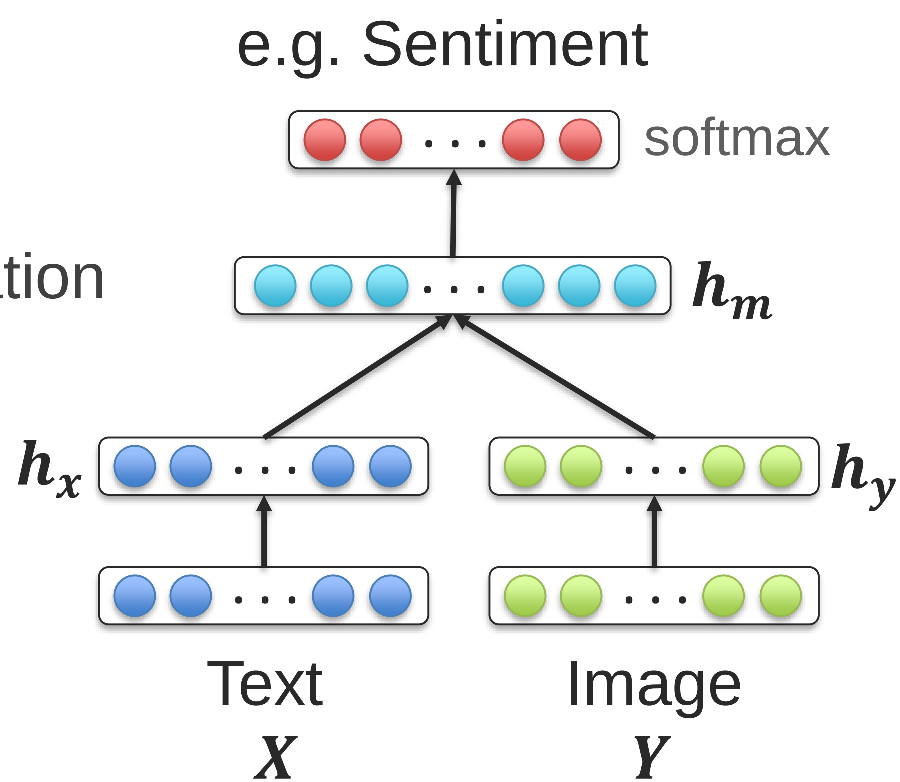
Multimodal Encoder-Decoder

- 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?



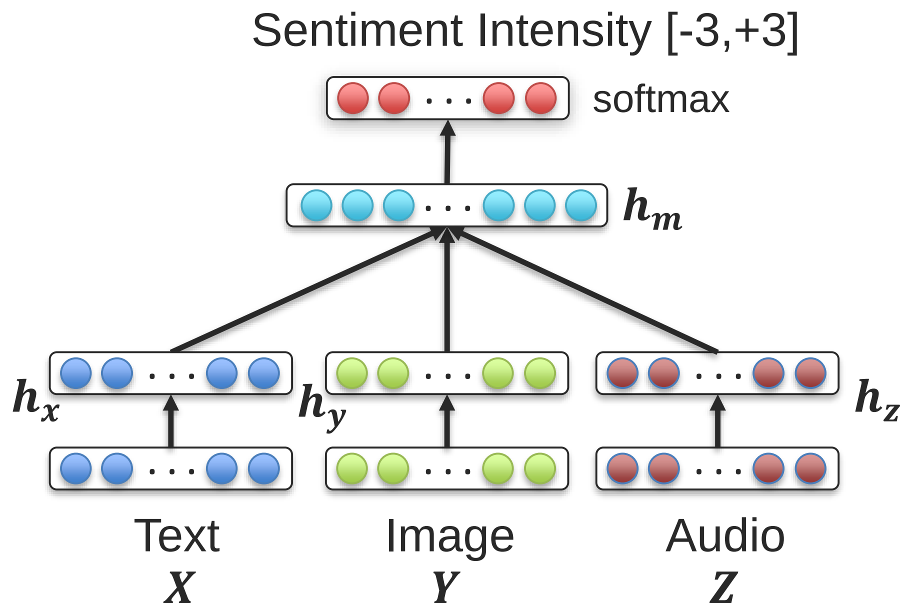
Multimodal Sentiment Analysis

MOSI dataset (Zadeh et al, 2016)

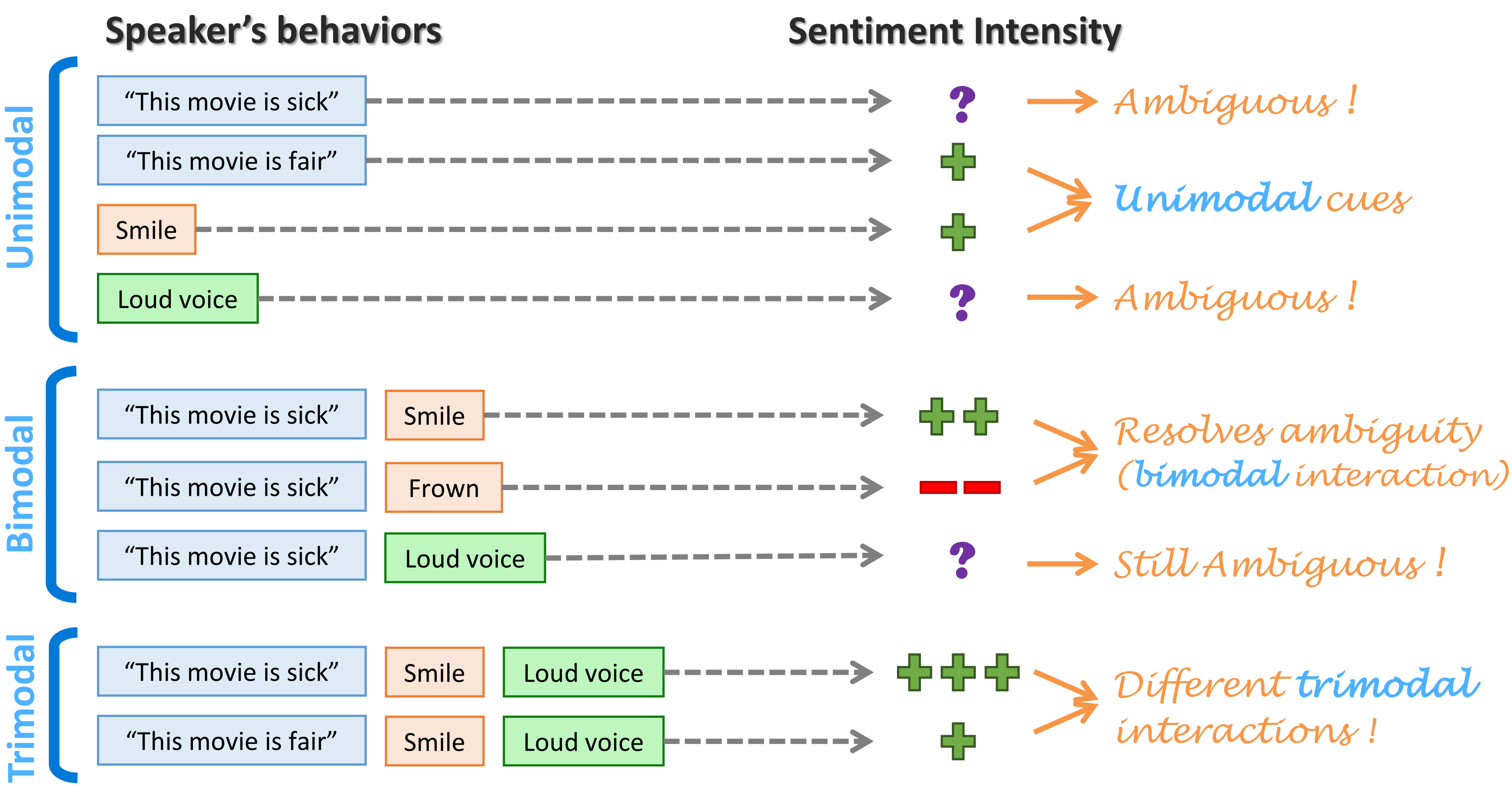
- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

Multimodal joint representation:

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



Unimodal, Bimodal and Trimodal Interactions

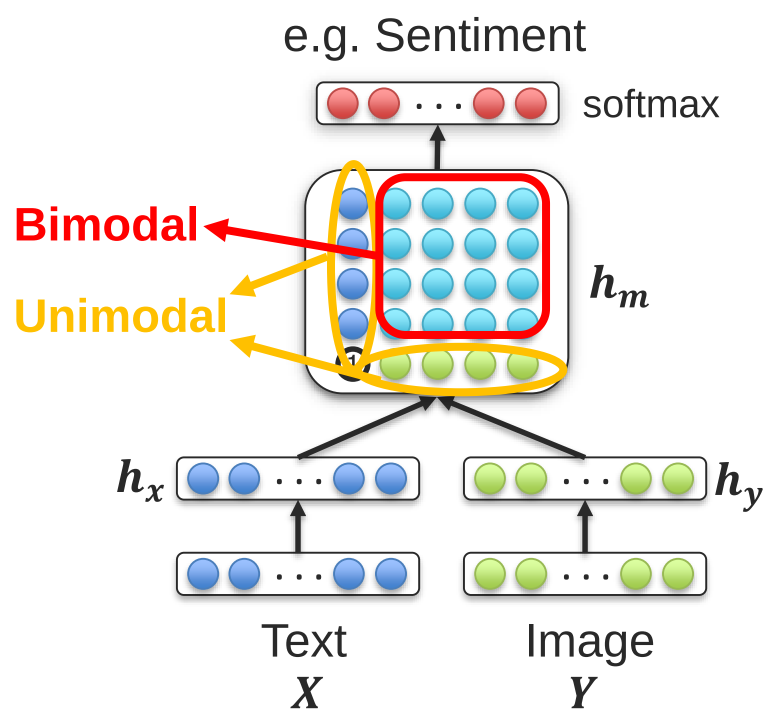


Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

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

Important !

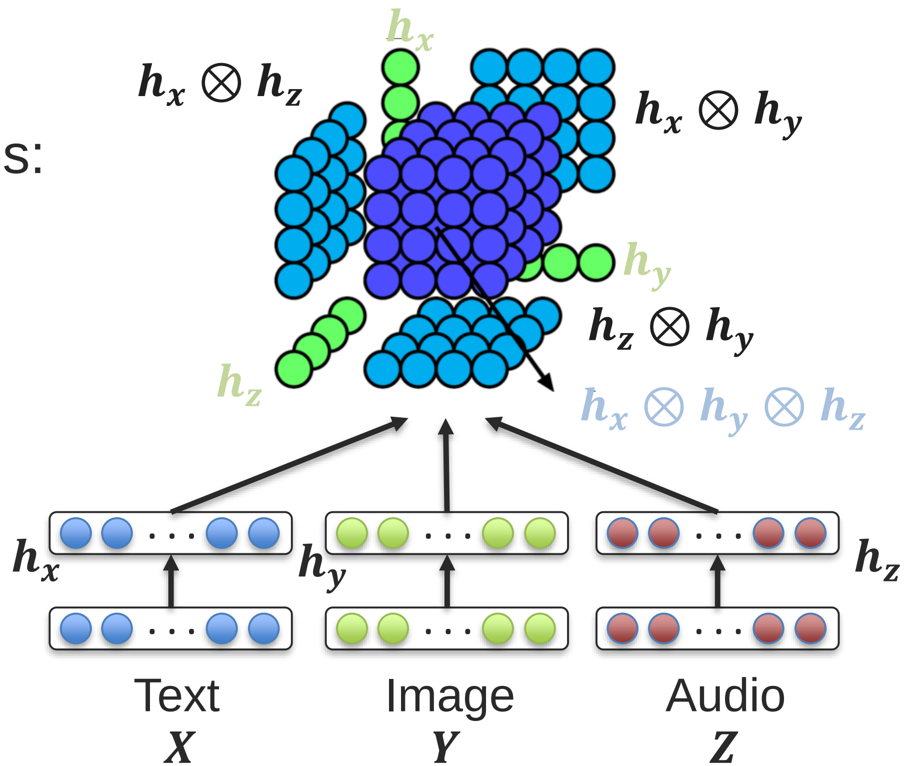


[Zadeh, Jones and Morency, **EMNLP 2017**]

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 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}	$\uparrow$ 4.0	$\uparrow$ 2.7	$\uparrow$ 6.7	$\downarrow$ 0.23	$\uparrow$ 0.17

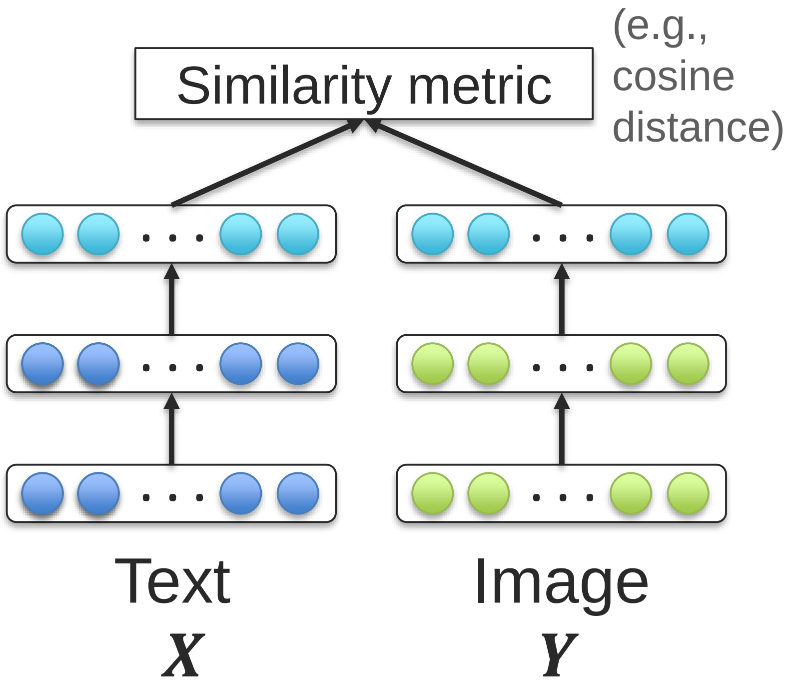
Improvement over State-Of-The-Art

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

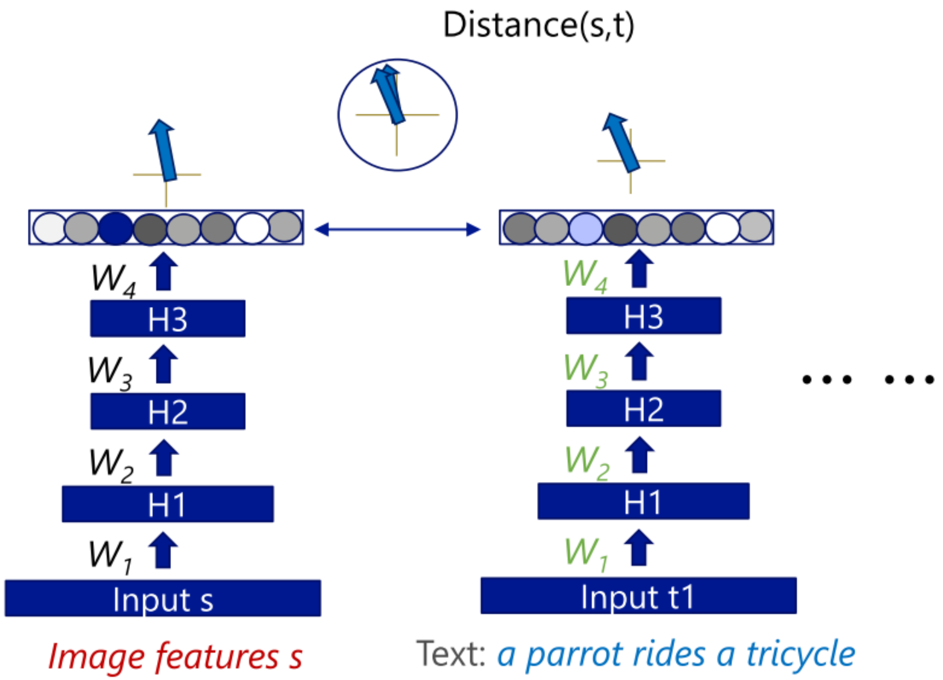
Coordinated Multimodal Representations

Coordinated Multimodal Representations

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

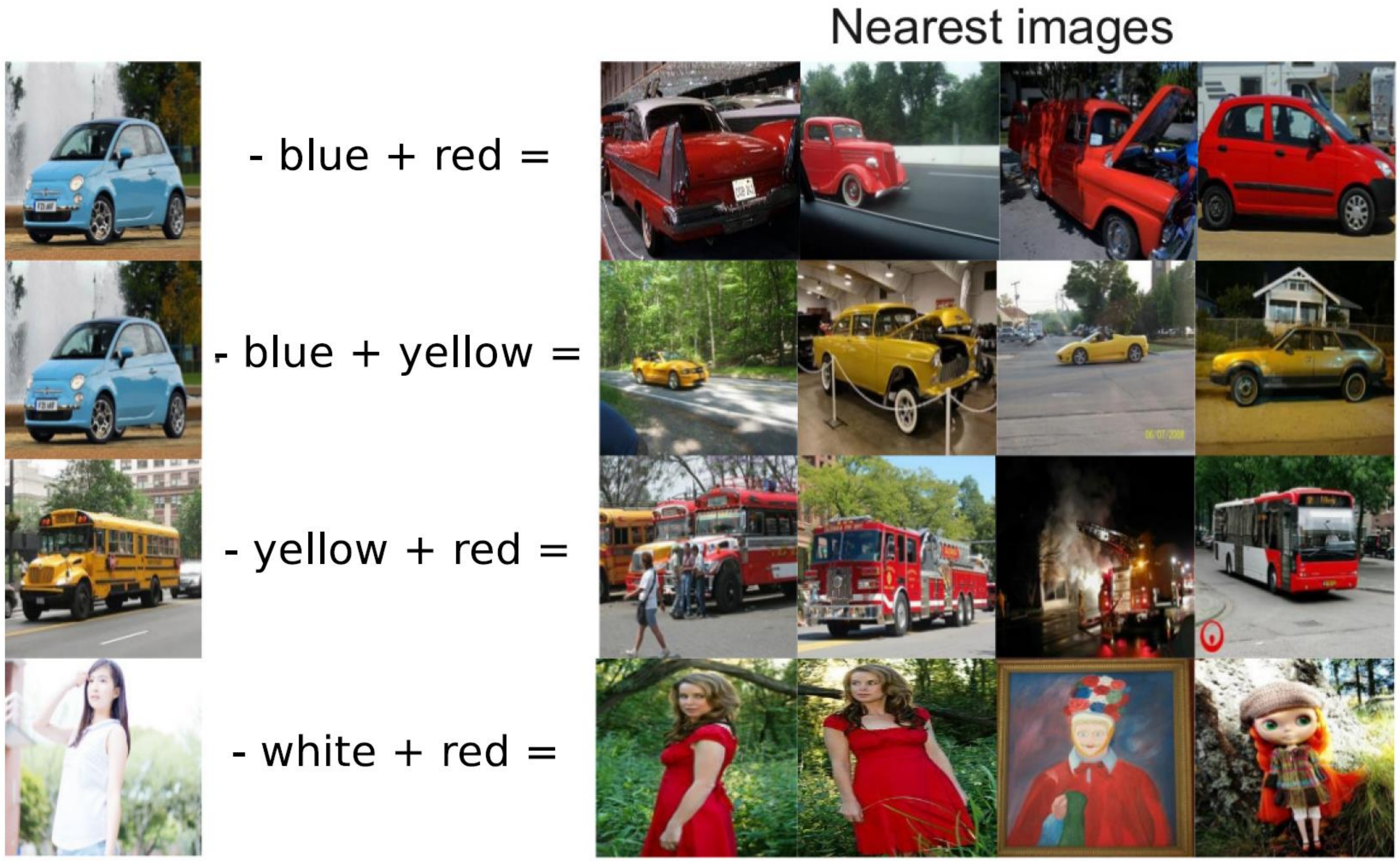


Coordinated Multimodal Embeddings



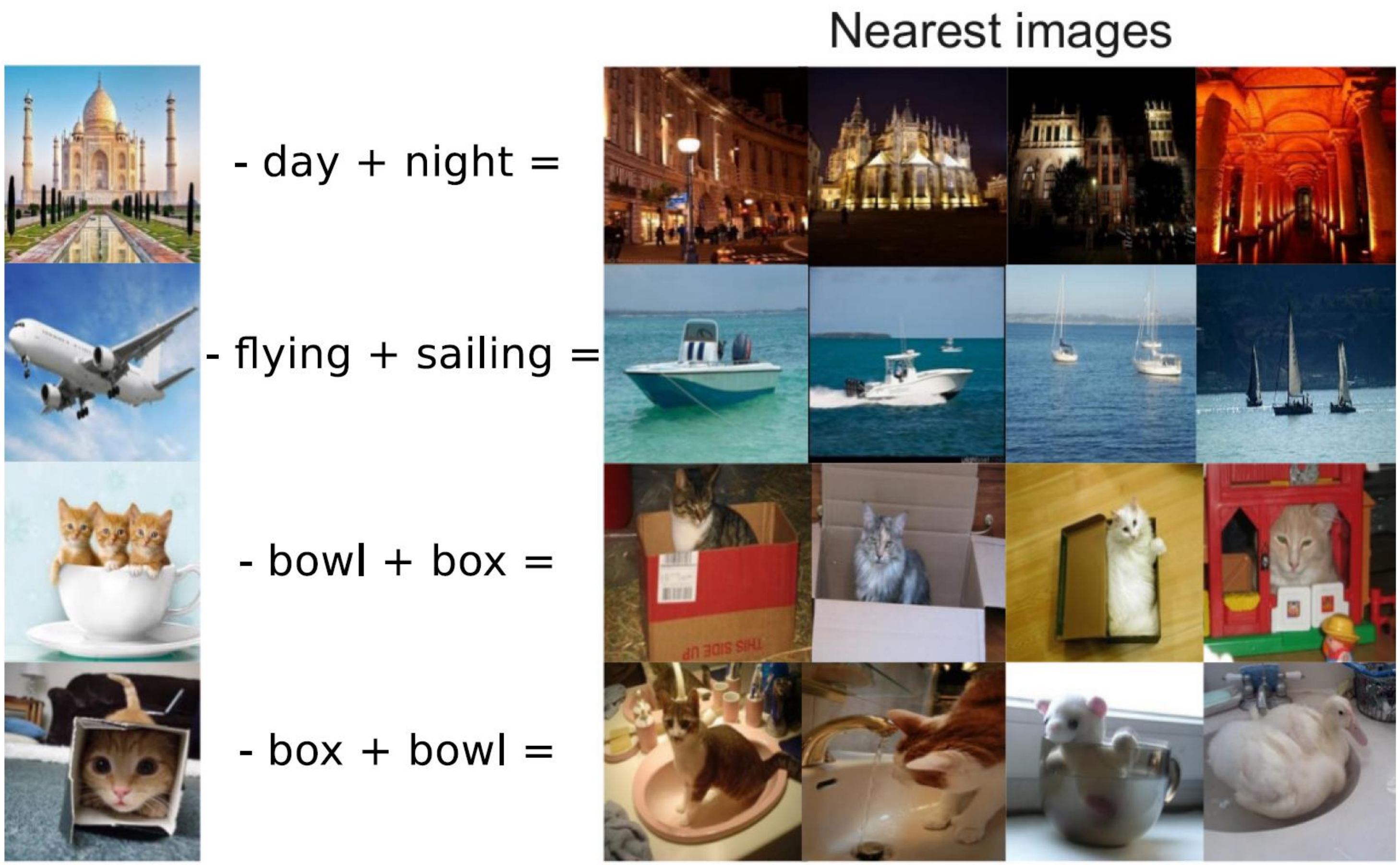
[Huang et al., Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, 2013]

Multimodal Vector Space Arithmetic



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

Multimodal Vector Space Arithmetic



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

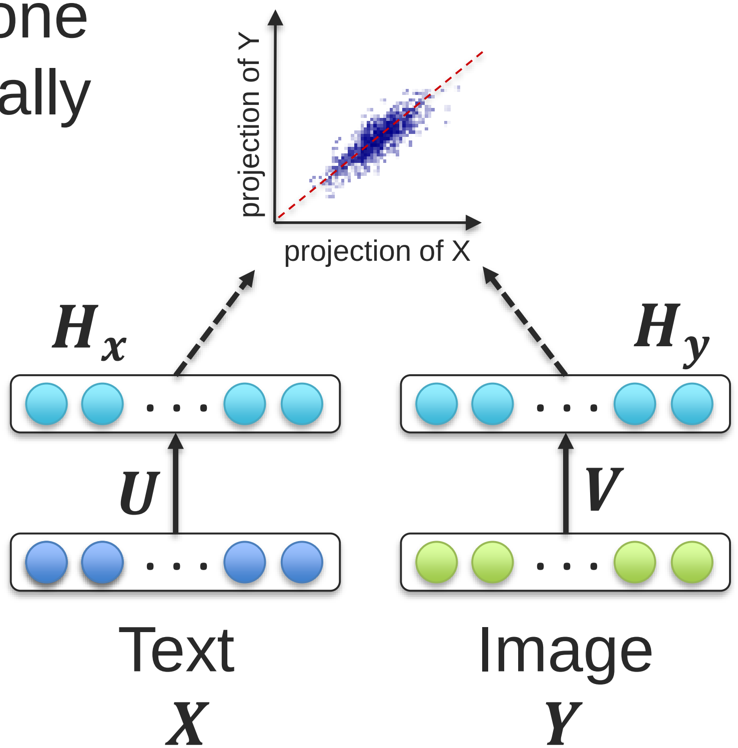
Canonical Correlation Analysis

"canonical": reduced to the simplest or clearest schema possible

1. Learn two linear projections, one for each view, that are maximally correlated:

$$(\boldsymbol{u}^*, \boldsymbol{v}^*) = \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr(\boldsymbol{H_x}, \boldsymbol{H_y})$$

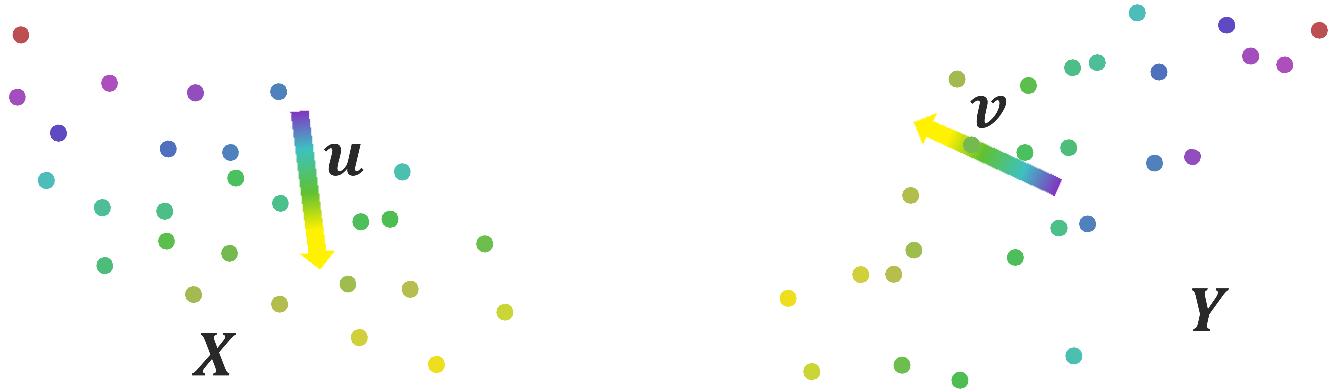
$$= \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr(\boldsymbol{u^T X}, \boldsymbol{v^T Y})$$



Correlated Projection

1 Learn two linear projections, one for each view, that are maximally correlated:

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



Two views $\boldsymbol{X}, \boldsymbol{Y}$ where same instances have the same color

Canonical Correlation Analysis

We want to learn multiple projection pairs $\left(\boldsymbol{u}_{(i)}\boldsymbol{X}, \boldsymbol{v}_{(i)}\boldsymbol{Y}\right)$:

$$\left(\boldsymbol{u}_{(i)}^{*}, \boldsymbol{v}_{(i)}^{*}\right) = \underset{\boldsymbol{u}_{(i)}, \boldsymbol{v}_{(i)}}{\operatorname{argmax}}\, corr\left(\boldsymbol{u}_{(i)}^{\boldsymbol{T}}\boldsymbol{X}, \boldsymbol{v}_{(i)}^{\boldsymbol{T}}\boldsymbol{Y}\right) \qquad \approx \boldsymbol{u}_{(i)}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{v}_{(i)}$$

2 We want these multiple projection pairs to be orthogonal ("canonical") to each other:

$$\boldsymbol{u}_{(i)}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{v}_{(j)} = \boldsymbol{u}_{(j)}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{v}_{(i)} = \boldsymbol{0} \qquad \text{for } i \neq j$$

$$\boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{V} = tr(\boldsymbol{U}\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{V}) \qquad \text{where } \boldsymbol{U} = [\boldsymbol{u}_{(1)}, \boldsymbol{u}_{(2)}, \ldots, \boldsymbol{u}_{(k)}]$$

$$\text{and } \boldsymbol{V} = [\boldsymbol{v}_{(1)}, \boldsymbol{v}_{(2)}, \ldots, \boldsymbol{v}_{(k)}]$$

Canonical Correlation Analysis

3 Since this objective function is invariant to scaling, we can constraint the projections to have unit variance:

$$\boldsymbol{U}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{XX}} \boldsymbol{U} = \boldsymbol{I} \qquad \boldsymbol{V}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{YY}} \boldsymbol{V} = \boldsymbol{I}$$

Canonical Correlation Analysis:

maximize: $tr(\boldsymbol{U}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{XY}} \boldsymbol{V})$

subject to: $\boldsymbol{U}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{YY}} \boldsymbol{U} = \boldsymbol{V}^{\boldsymbol{T}} \boldsymbol{\Sigma}_{\boldsymbol{YY}} \boldsymbol{V} = \boldsymbol{I}$

Canonical Correlation Analysis

maximize: $tr(\boldsymbol{U}^T\boldsymbol{\Sigma}_{XY}\boldsymbol{V})$

subject to: $\boldsymbol{U}^T\boldsymbol{\Sigma}_{YY}\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{\Sigma}_{YY}\boldsymbol{V} = \boldsymbol{I}$

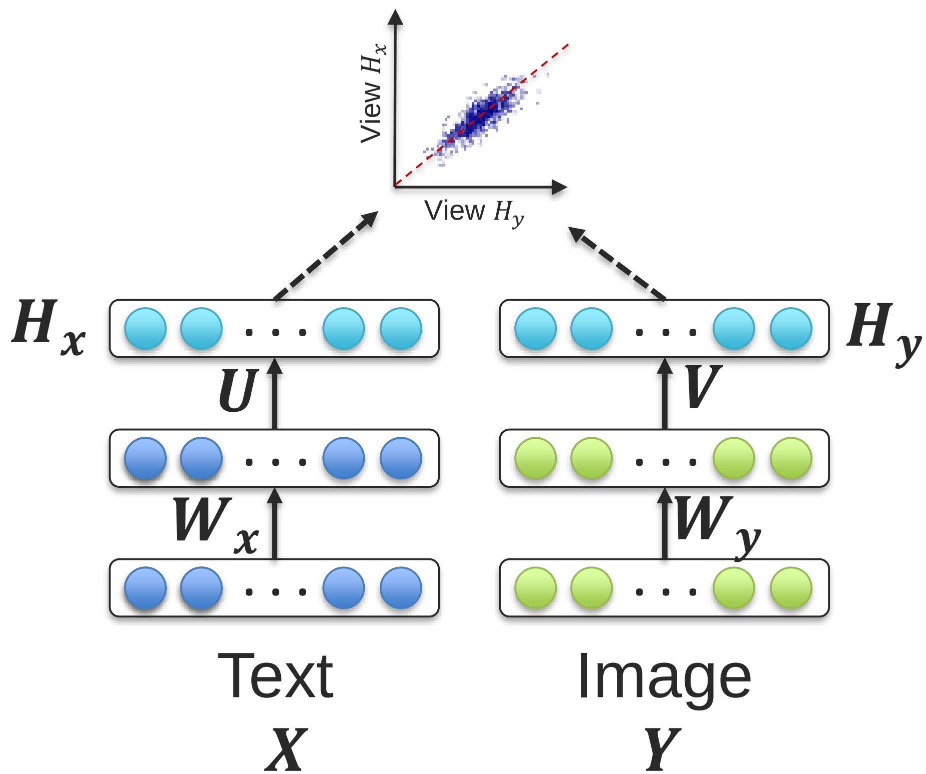
$$\Sigma = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{XX} & \boldsymbol{\Sigma}_{YX} \\ \hline \boldsymbol{\Sigma}_{XY} & \boldsymbol{\Sigma}_{YY} \end{array}\right] \overset{\boldsymbol{U},\boldsymbol{V}}{\Longrightarrow} \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & \lambda_1 & 0 & 0 \\ 0 & 1 & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 1 & 0 & 0 & \lambda_3 \\ \hline \lambda_1 & 0 & 0 & 1 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_3 & 0 & 0 & 1 \end{array}\right]$$

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

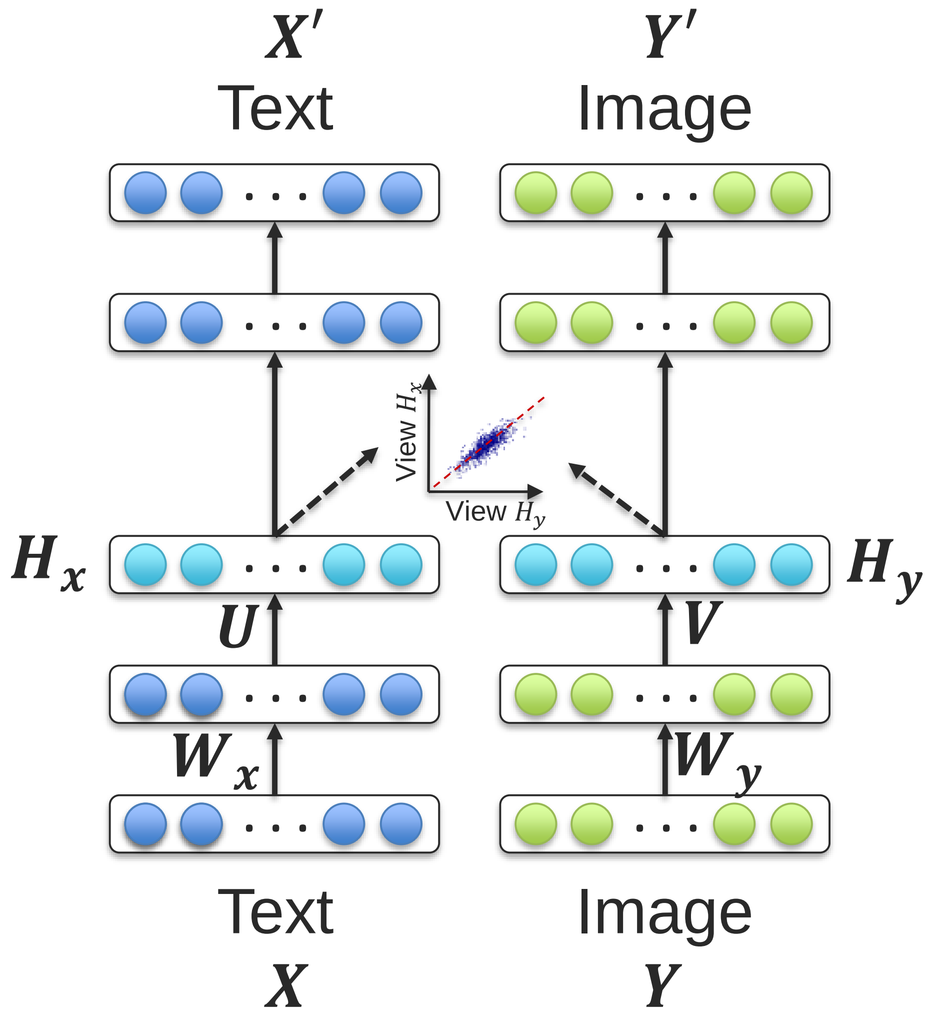


Andrew et al., ICML 2013

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

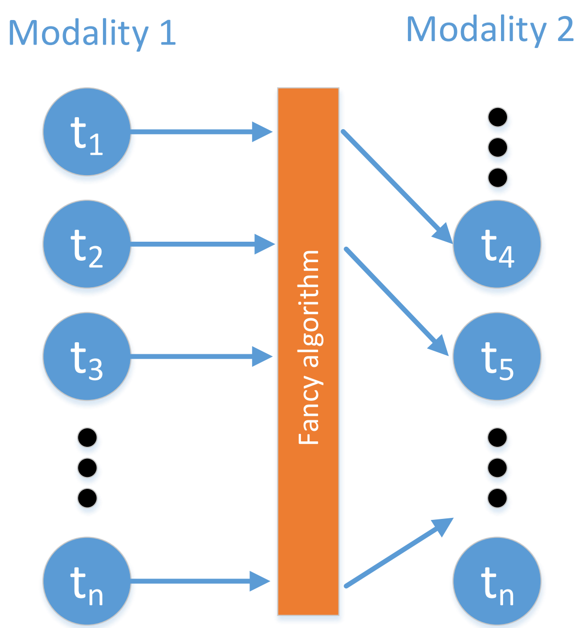


Wang et al., ICML 2015

Explicit alignment

Core Challenge: Alignment

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



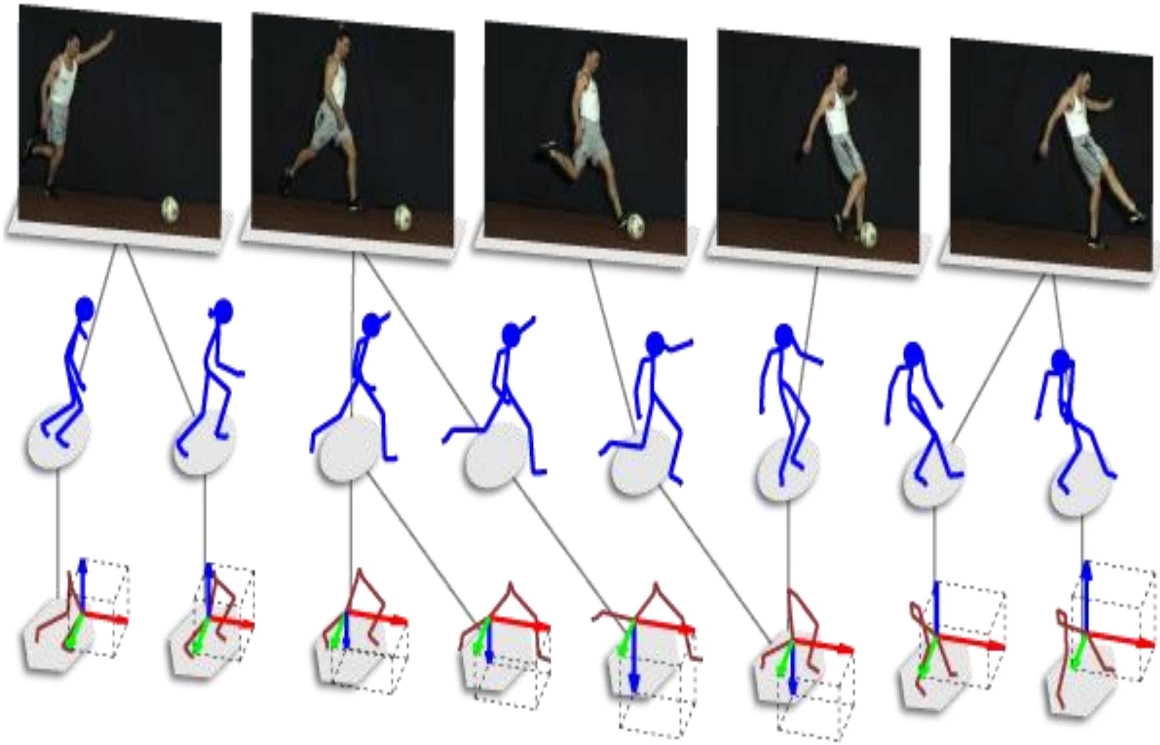
A Explicit Alignment

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



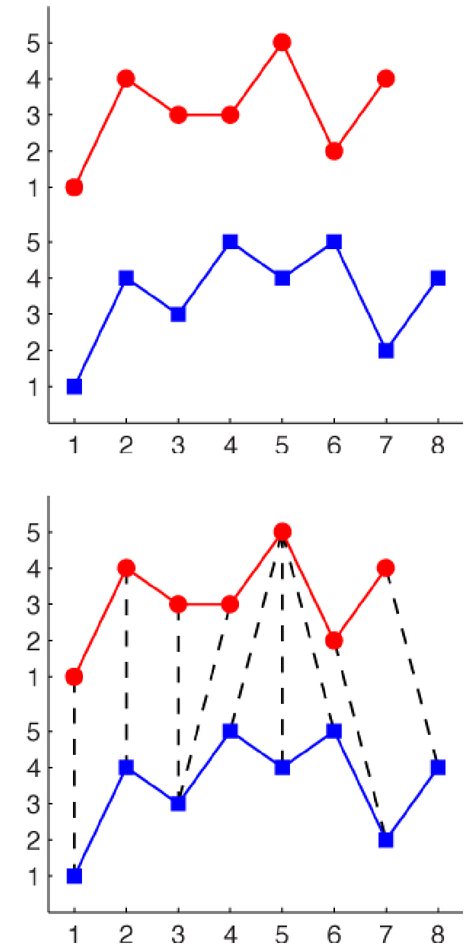
Applications:
- Re-aligning asynchronous data
- Finding similar data across modalities (we can estimate the aligned cost)
- Event reconstruction from multiple sources

Let’s start unimodal – Dynamic Time Warping

- We have two unaligned temporal unimodal signals
  - $\mathbf{X} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_{n_x}\right] \in \mathbb{R}^{d \times n_x}$
  - $\mathbf{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_{n_y}\right] \in \mathbb{R}^{d \times n_y}$
- Find set of indices to minimize the alignment difference:

$$L(\boldsymbol{p}_t^x, \boldsymbol{p}_t^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{\boldsymbol{p}_t^x} - \boldsymbol{y}_{\boldsymbol{p}_t^y} \right\|_2^2$$

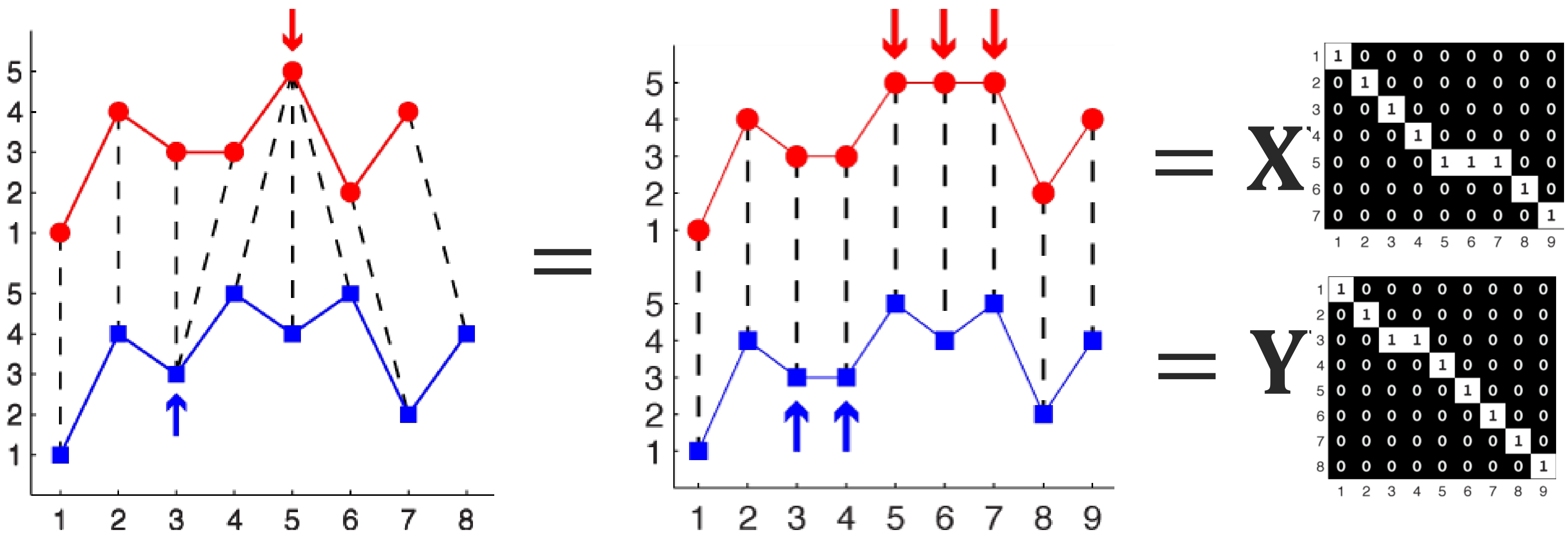
- Where $\boldsymbol{p}_t^x$ and $\boldsymbol{p}_t^y$ are index vectors of same length
- Finding these indices is called Dynamic Time Warping



DTW alternative formulation

$$L(\boldsymbol{p}_t^x, \boldsymbol{p}_t^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{\boldsymbol{p}_t^x} - \boldsymbol{y}_{\boldsymbol{p}_t^y} \right\|_2^2$$

Replication doesn't change the objective!



Alternative objective:

$$L(\boldsymbol{W}_x, \boldsymbol{W}_y) = \left\| \boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y \right\|_F^2$$

$\boldsymbol{X}, \boldsymbol{Y}$ – original signals (same #rows, possibly different #columns)

$\boldsymbol{W}_x, \boldsymbol{W}_y$ - alignment matrices

Frobenius norm $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j |a_{i,j}|^2$

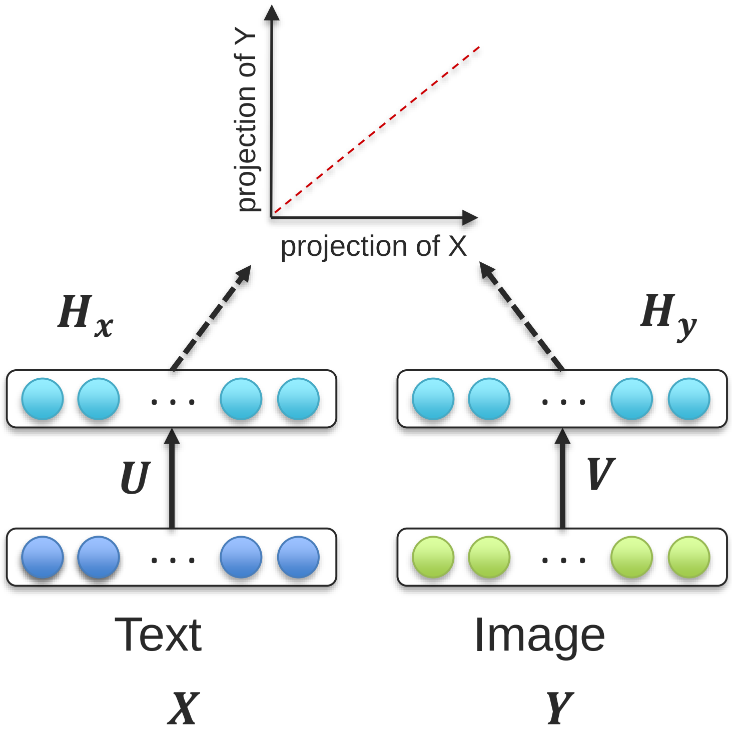
But how to handle multimodal data?

Canonical Correlation Analysis reminder

- When data is normalized it is actually equivalent to smallest RMSE reconstruction
- CCA loss can also be re-written as:

$$L(\boldsymbol{U},\boldsymbol{V}) = \|\mathbf{U}^T\mathbf{X} - \mathbf{V}^T\mathbf{Y}\|_F^2$$

subject to: $\boldsymbol{U}^T\boldsymbol{\Sigma}_{\boldsymbol{YY}}\boldsymbol{U} = \boldsymbol{V}^T\boldsymbol{\Sigma}_{\boldsymbol{YY}}\boldsymbol{V} = \boldsymbol{I}$



Canonical Time Warping

- Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

$$L(\boldsymbol{U},\boldsymbol{V},\boldsymbol{W}_x,\boldsymbol{W}_y) = \left\|\mathbf{U}^T\mathbf{X}\mathbf{W}_\mathbf{x} - \mathbf{V}^T\mathbf{Y}\mathbf{W}_\mathbf{y}\right\|_F^2$$

- Allows to align multi-modal or multi-view (same modality but from a different point of view)
- $\boldsymbol{W}_x, \boldsymbol{W}_y$ – temporal alignment
- $\boldsymbol{U}, \boldsymbol{V}$ – cross-modal (spatial) alignment

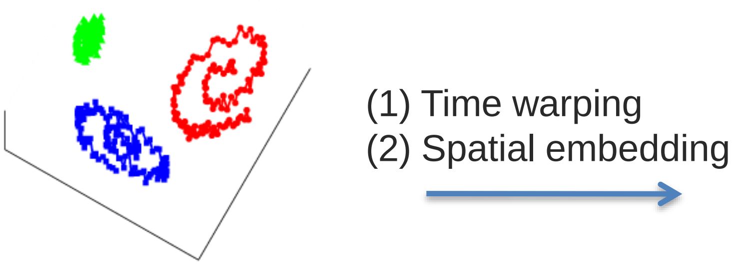
[Canonical Time Warping for Alignment of Human Behavior, Zhou and De la Tore, 2009]

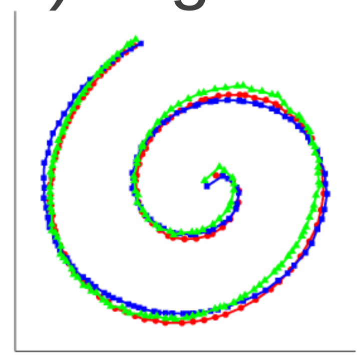
Generalized Time warping

- Generalize to multiple sequences all of different modality

$$L(\boldsymbol{U_i}, \boldsymbol{W_i}) = \sum_{i=1} \sum_{j=1} \left\| \mathbf{U}_i^T \mathbf{X}_{\mathrm{i}} \mathbf{W}_{\mathrm{i}} - \mathbf{U}_j^T \mathbf{X}_{\mathrm{j}} \mathbf{W}_j \right\|_F^2$$

- $\boldsymbol{W_i}$ – set of temporal alignments
- $\boldsymbol{U_i}$ – set of cross-modal (spatial) alignments

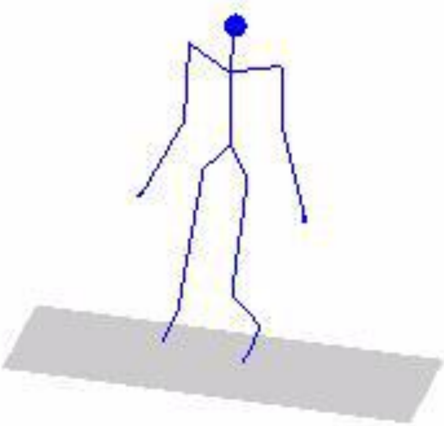




[Generalized Canonical Time Warping, Zhou and De la Tore, 2016, TPAMI]

Alignment examples (multimodal)

1/273



1/51

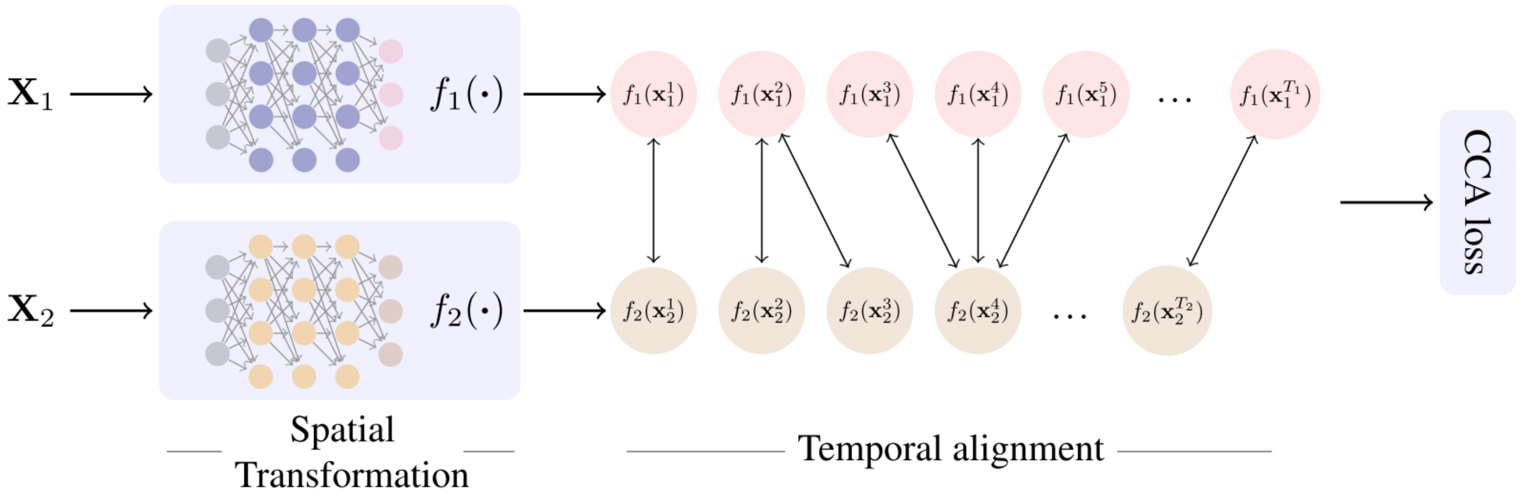
1/127

But how to model non-linear alignment functions?

Deep Canonical Time Warping

$$L(\theta_1, \theta_2, W_x, W_y) = \left\| f_{\theta_1}(\mathbf{X})\mathbf{W_x} - f_{\theta_1}(\mathbf{Y})\mathbf{W_y} \right\|_F^2$$

- Could be seen as generalization of DCCA and GTW



[Deep Canonical Time Warping, Trigeorgis et al., 2016, CVPR]

Implicit alignment

Machine Translation

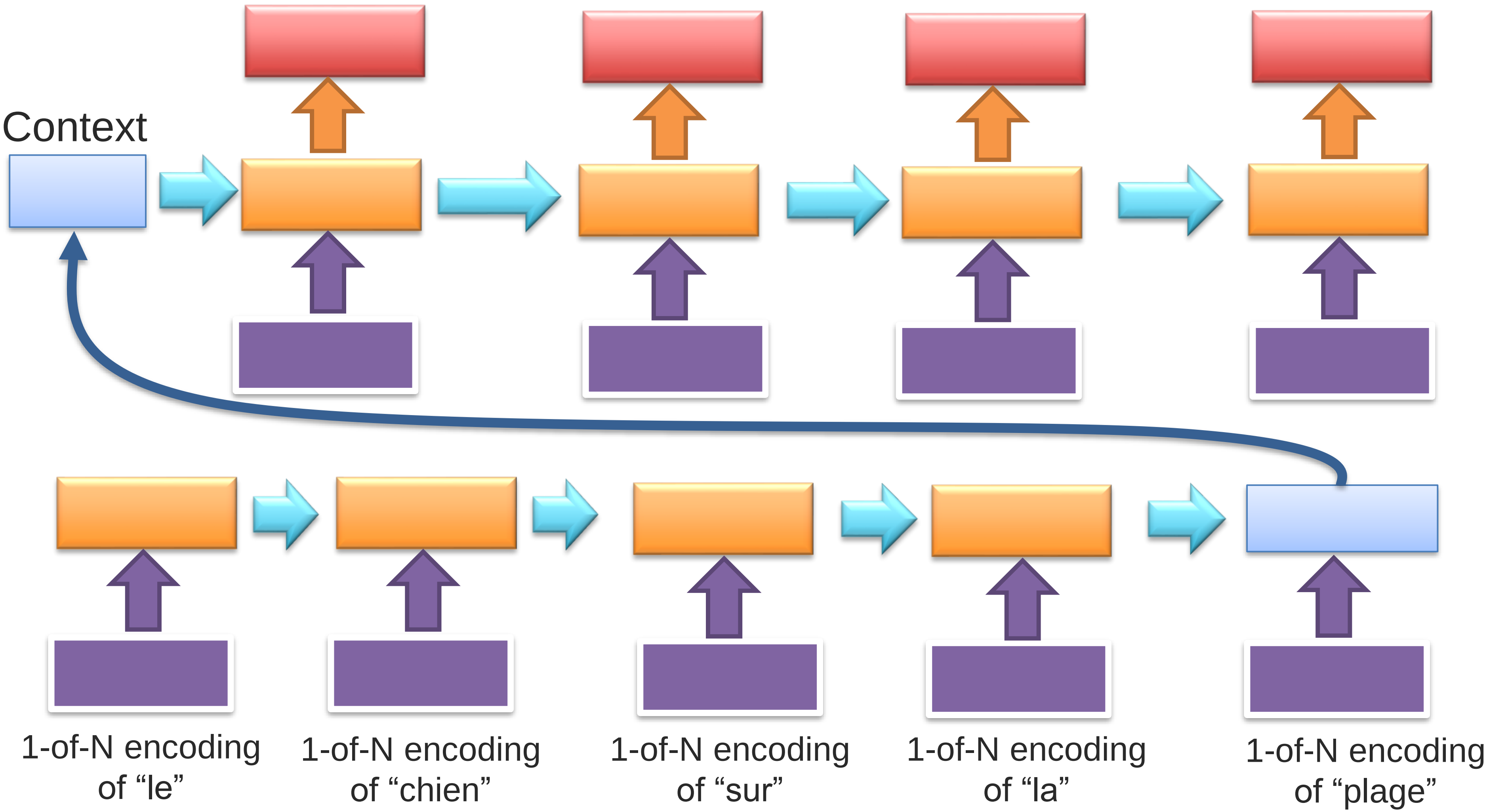
- Given a sentence in one language translate it to another

Dog on the beach → le chien sur la plage

- Not exactly multimodal task – but a good start! Each language can be seen almost as a modality.

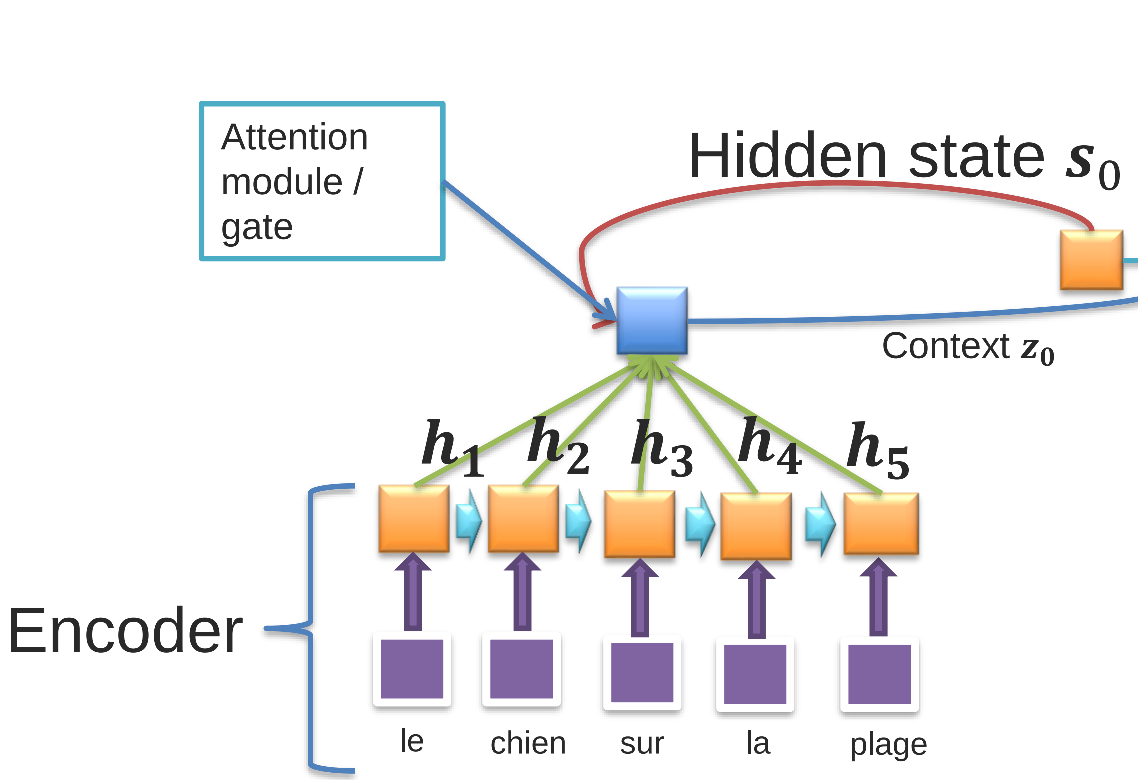
Encoder-Decoder Architecture for Machine Translation

[Cho et al., "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation", EMNLP 2014]



Attention Model for Machine Translation

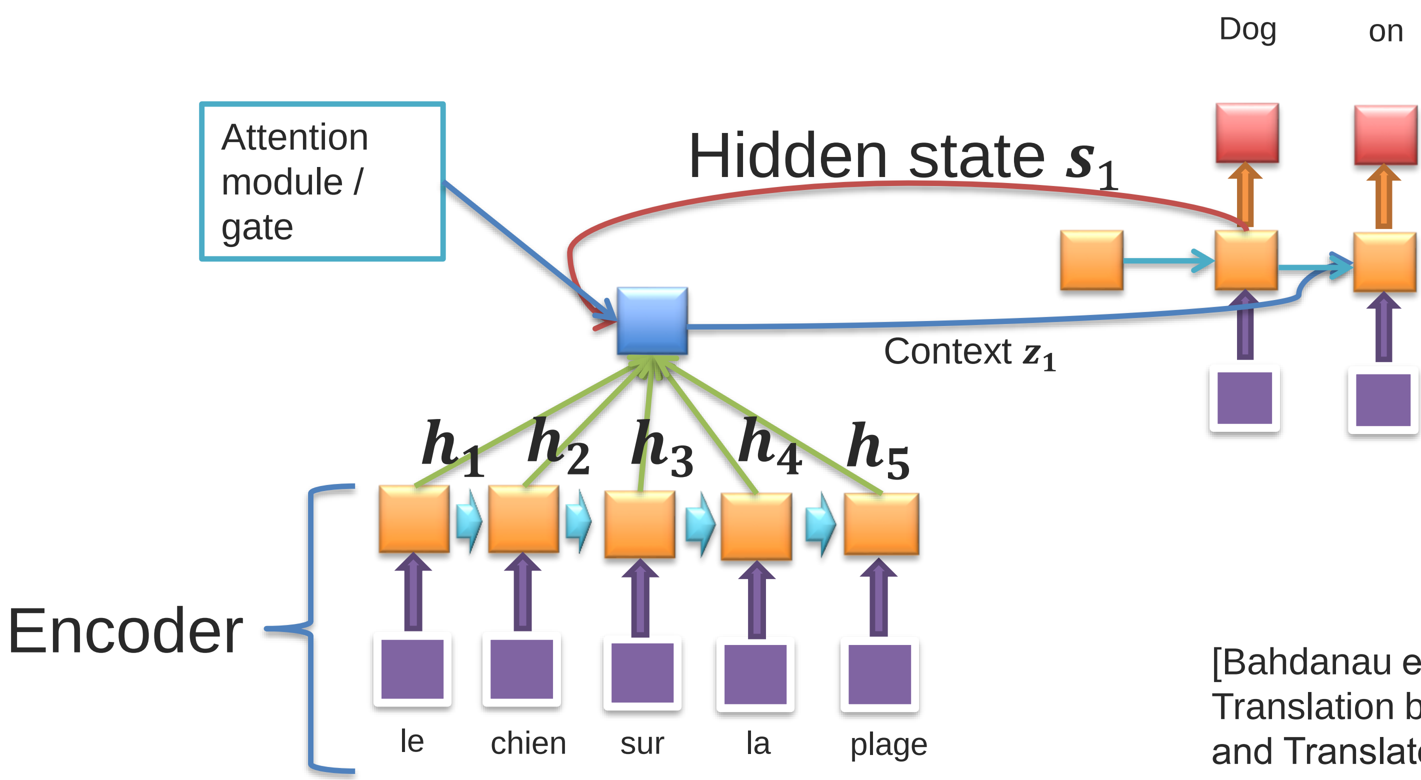
- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



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

Attention Model for Machine Translation

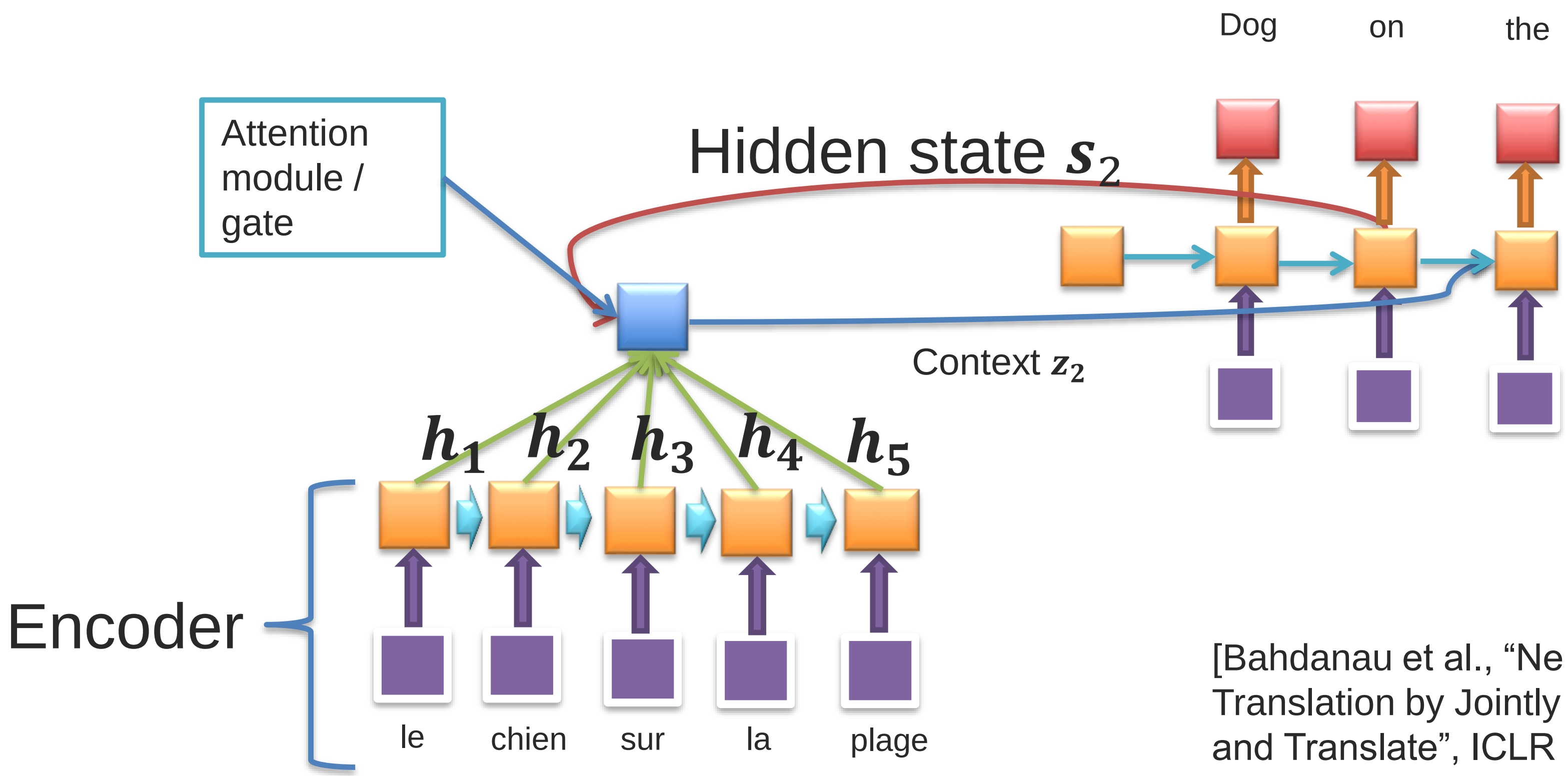
- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states



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

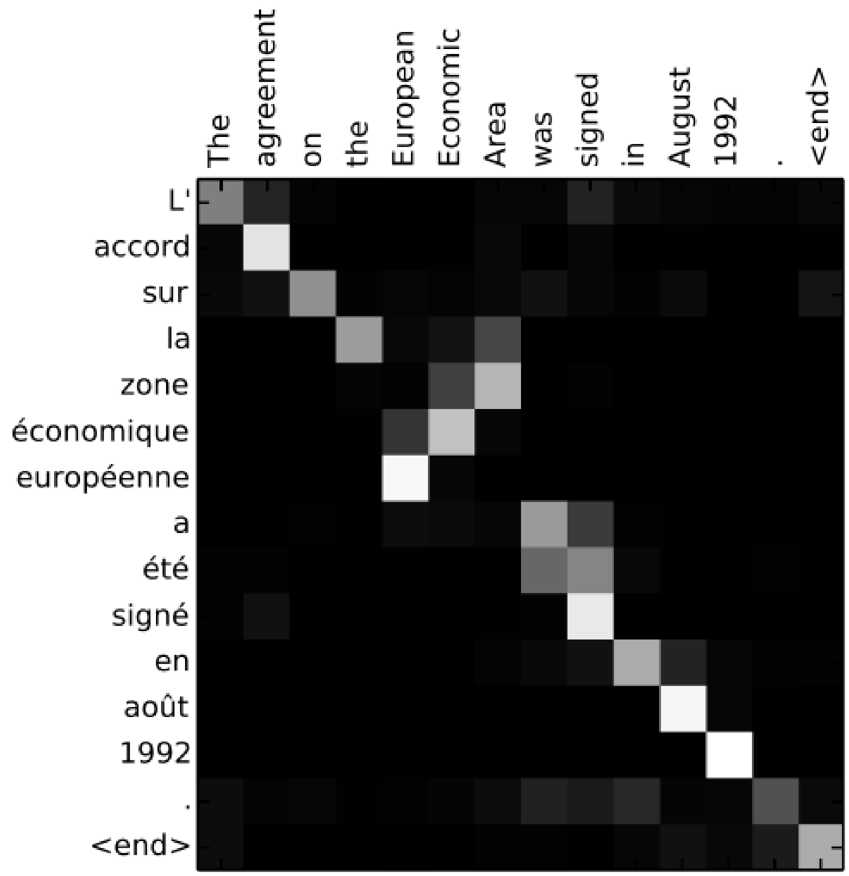
Attention Model for Machine Translation

- Before encoder would just take the final hidden state, now we actually care about the intermediate hidden states

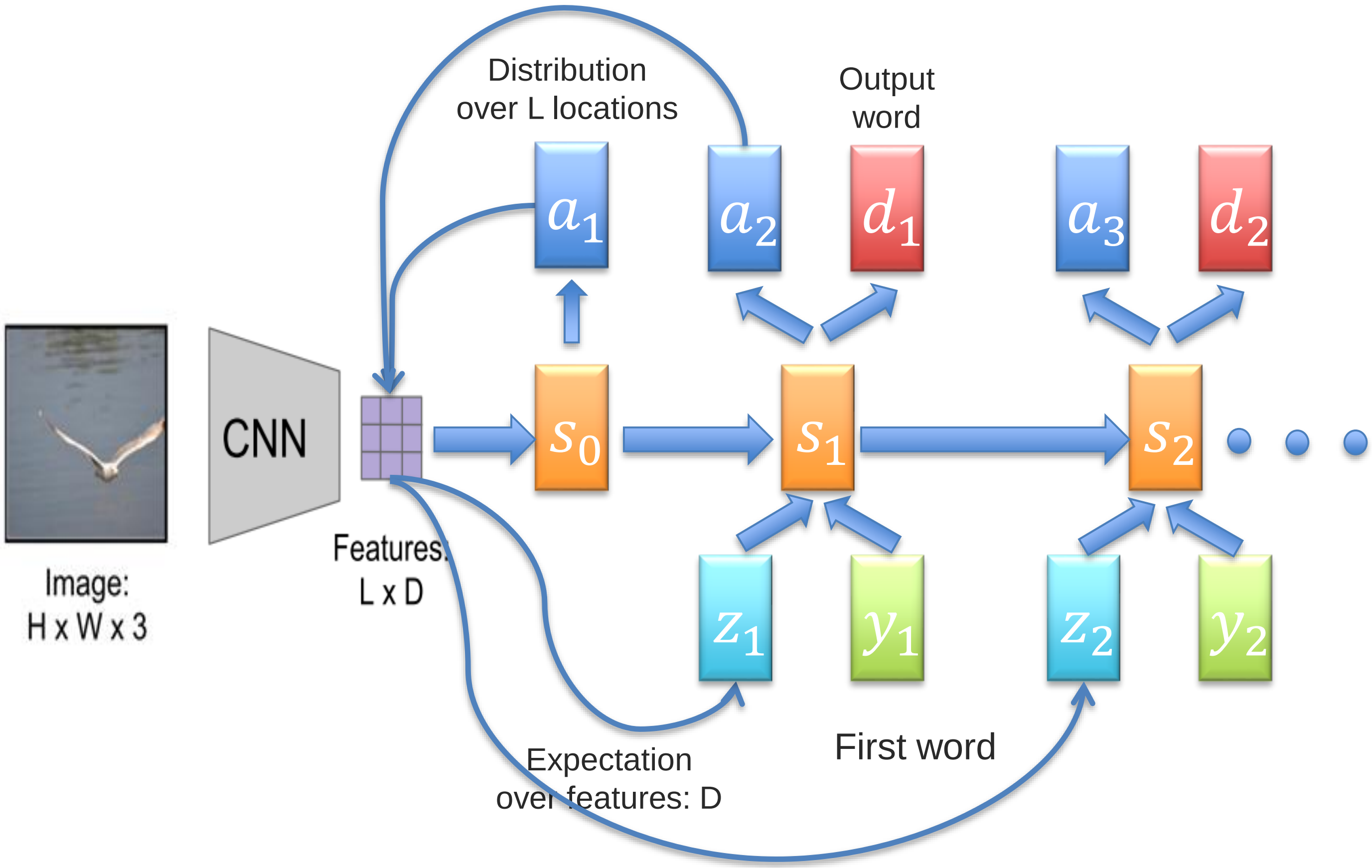


[Bahdanau et al., "Neural Machine Translation by Jointly Learning to Align and Translate", ICLR 2015]

Attention Model for Machine Translation



Attention Model for Image Captioning



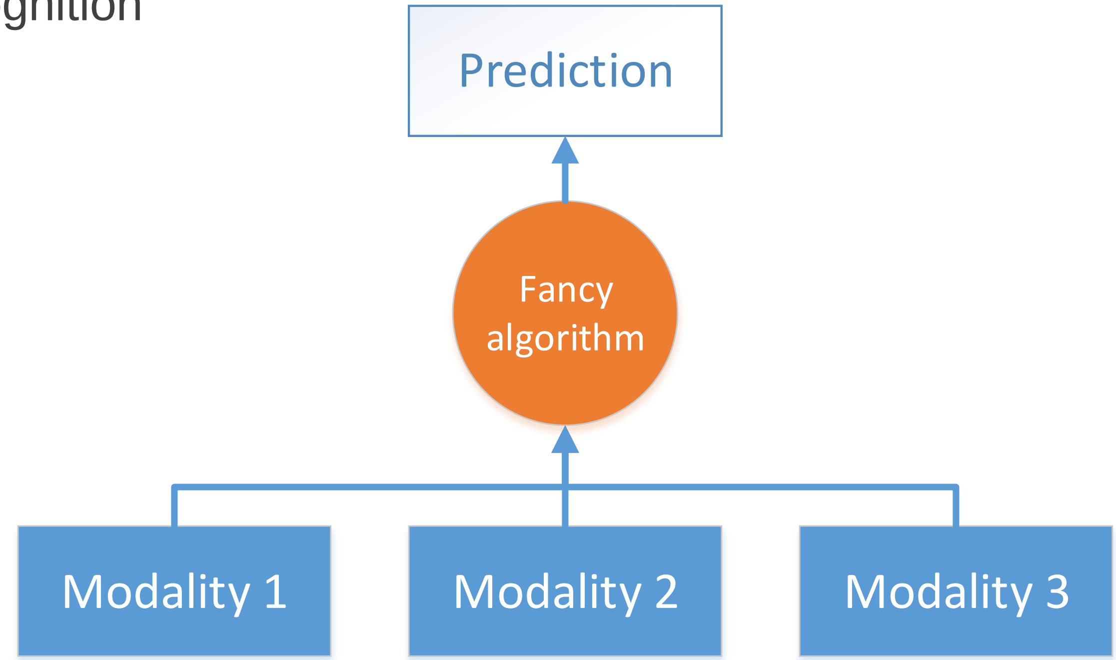
Attention Model for Image Captioning

Xu et.al., ICML 2015

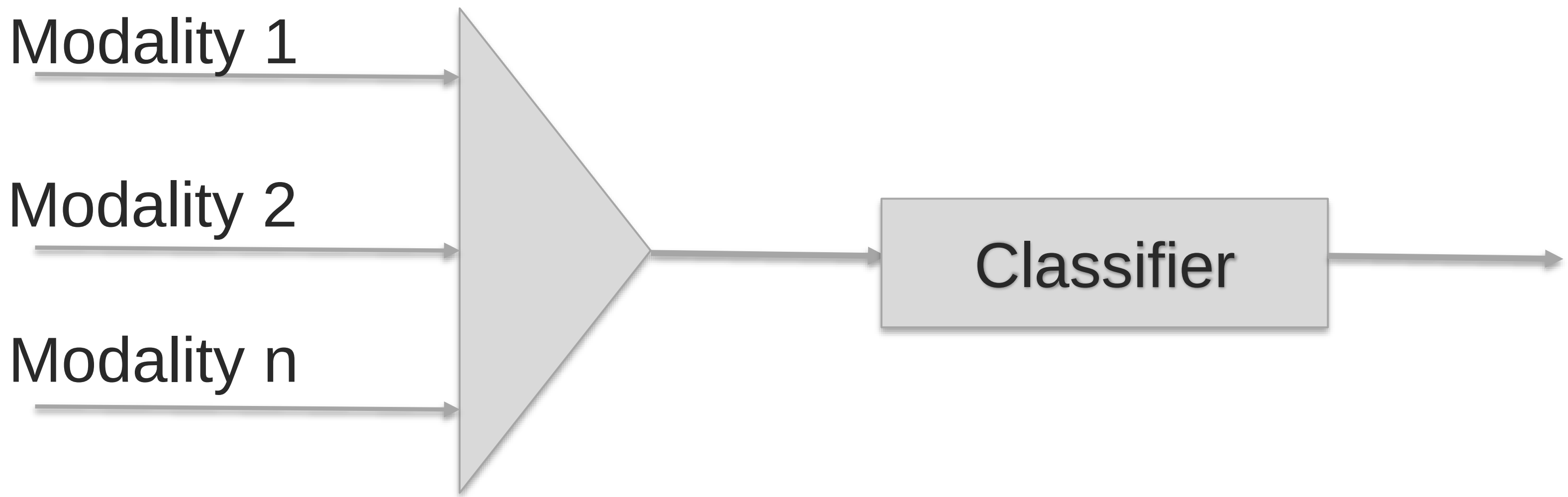
Multimodal Fusion

Multimodal 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

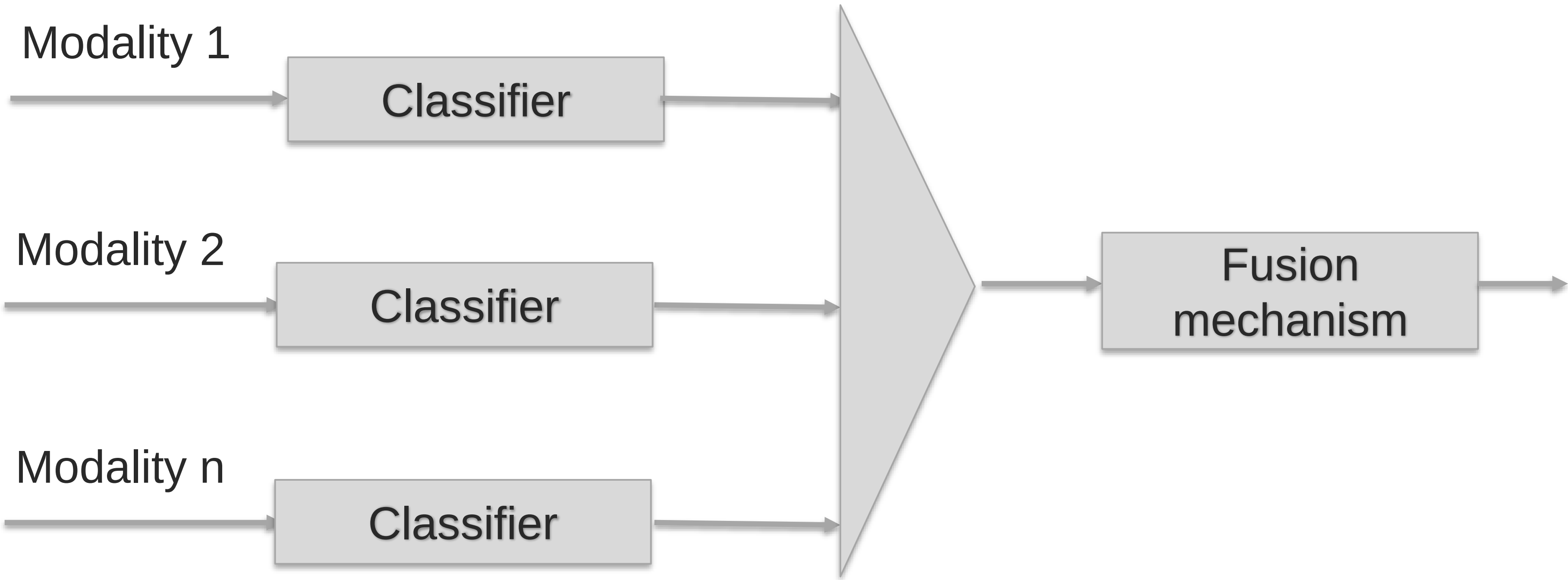


Model free approaches – early fusion



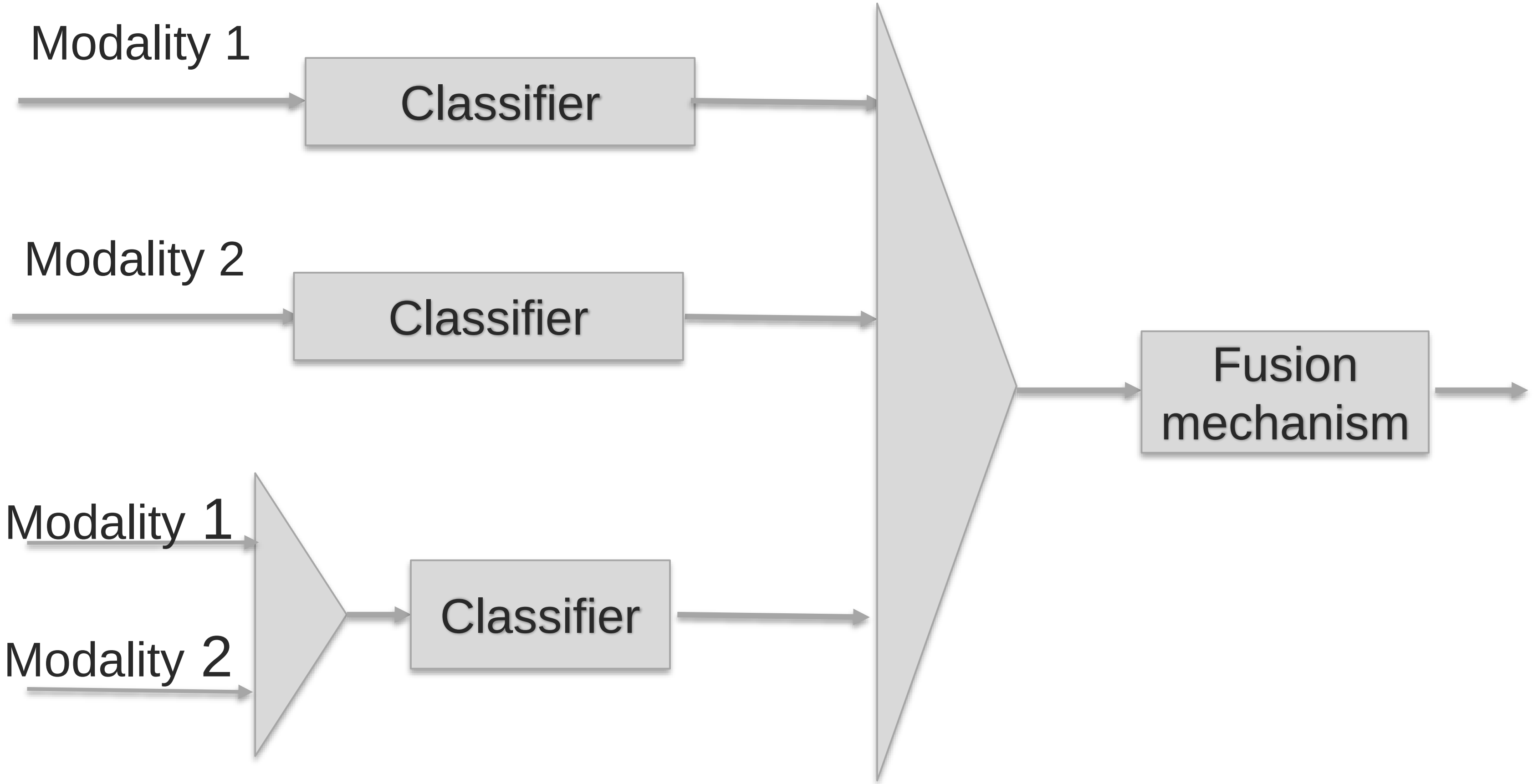
- Easy to implement – just concatenate the features
- Exploit dependencies between features
- Can end up very high dimensional
- More difficult to use if features have different framerates

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

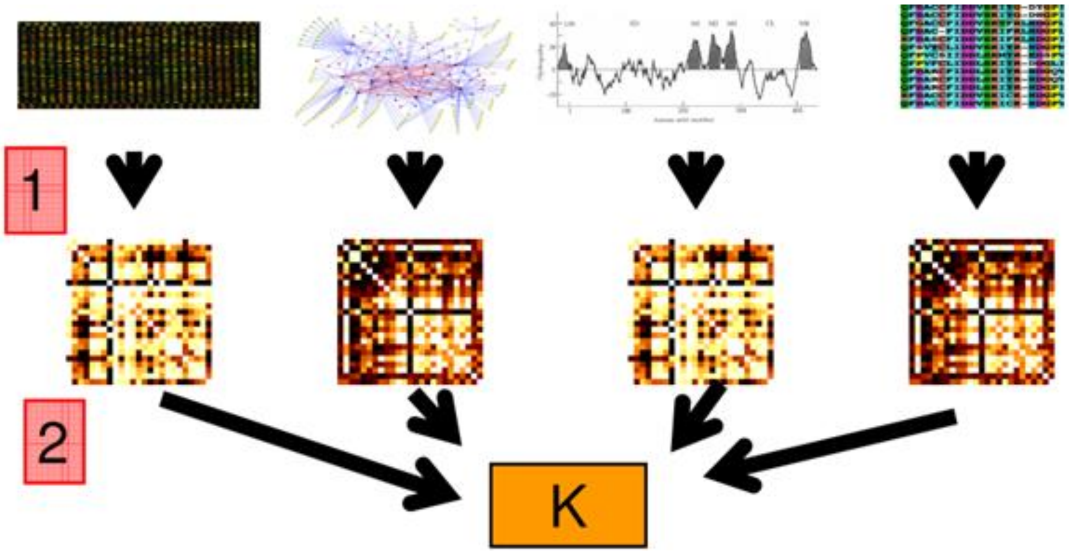
Model free approaches – hybrid fusion



- Combine benefits of both early and late fusion mechanisms

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

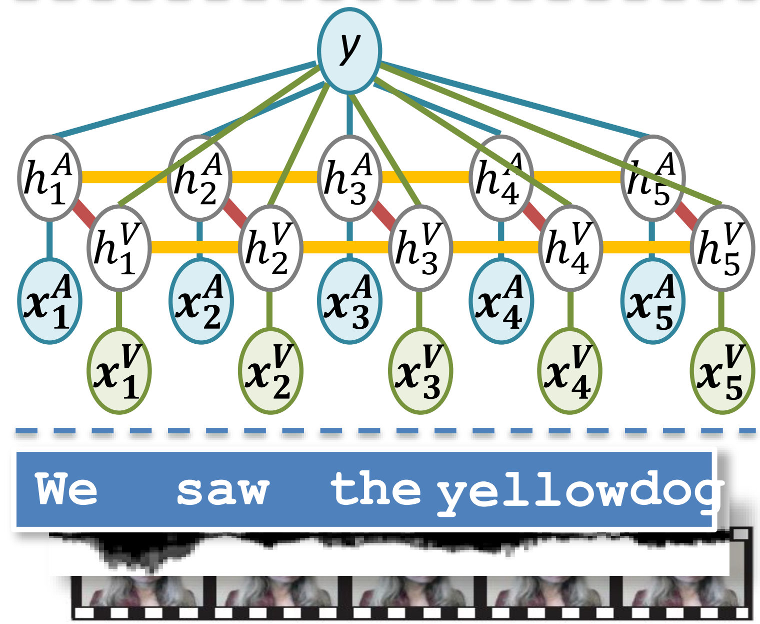
- Internal grouping of observations

Modality-***shared*** structure

- Interaction and synchrony

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

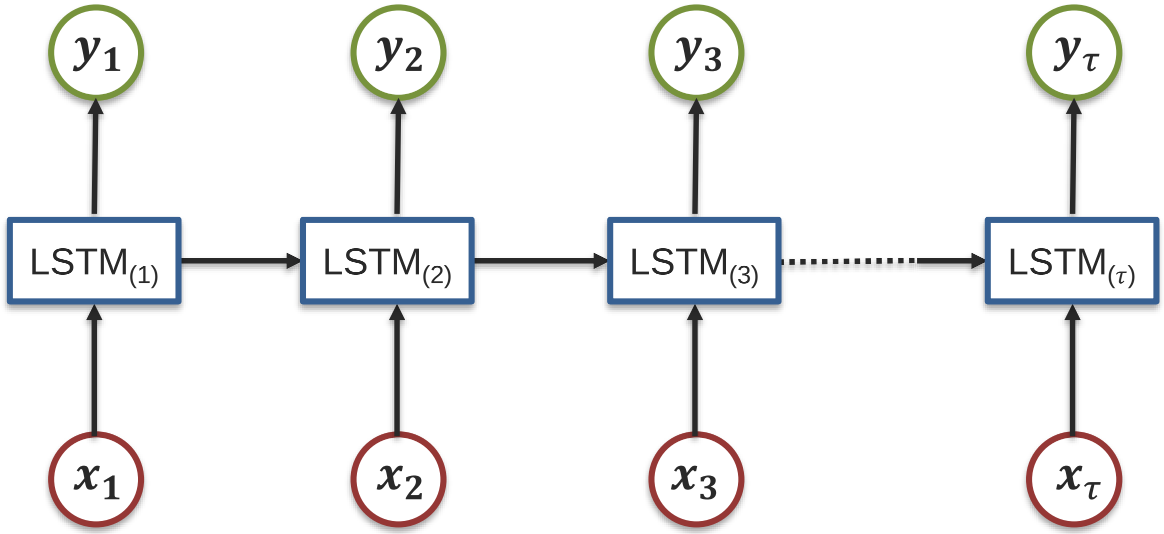
Multi-View
Hidden Conditional Random Field



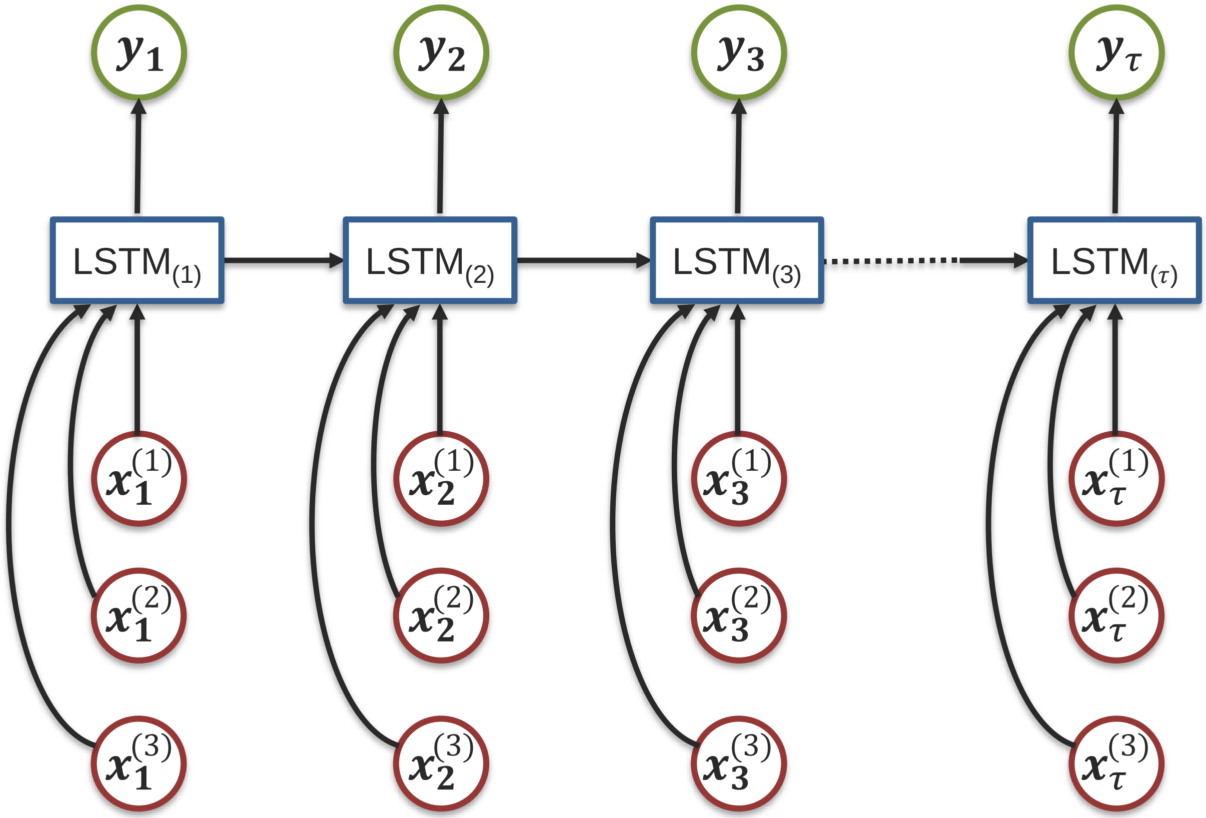
➢ Approximate inference using loopy-belief

[Song, Morency and Davis, CVPR 2012]

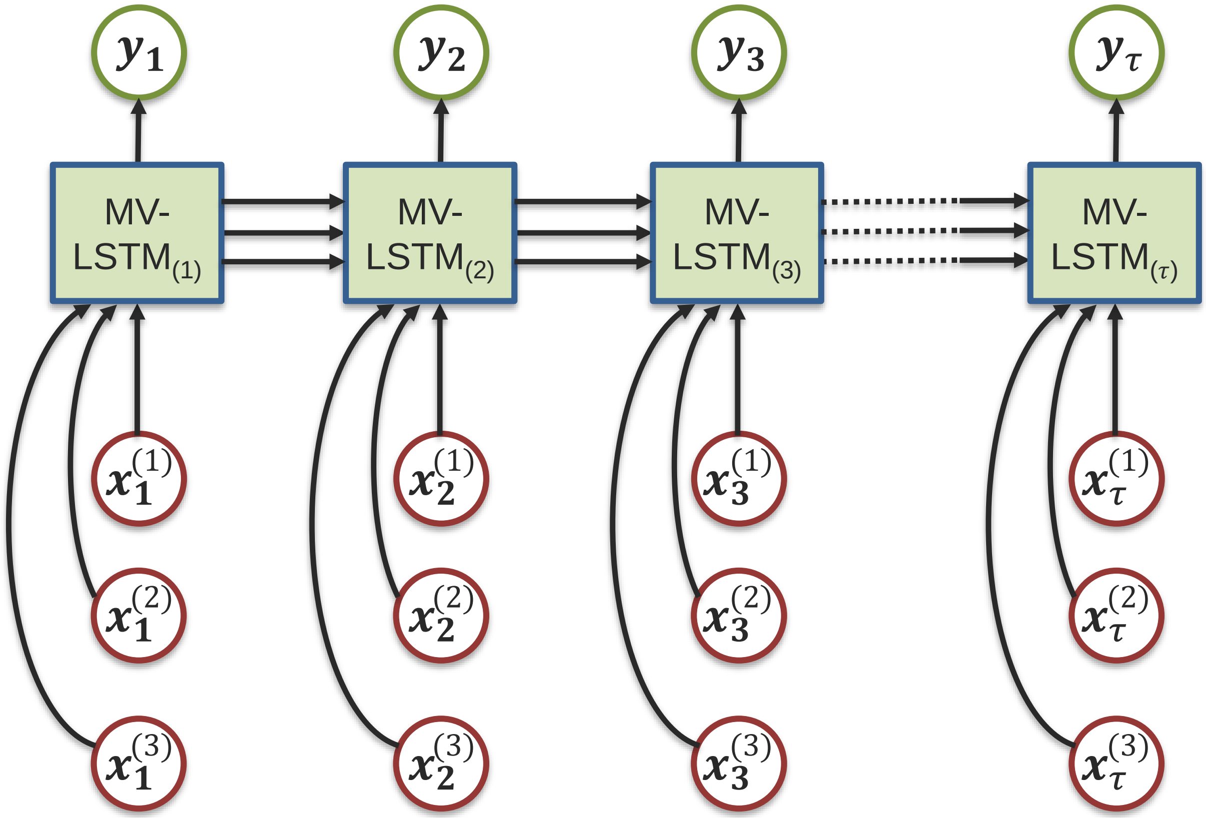
Sequence Modeling with LSTM



Multimodal Sequence Modeling – Early Fusion

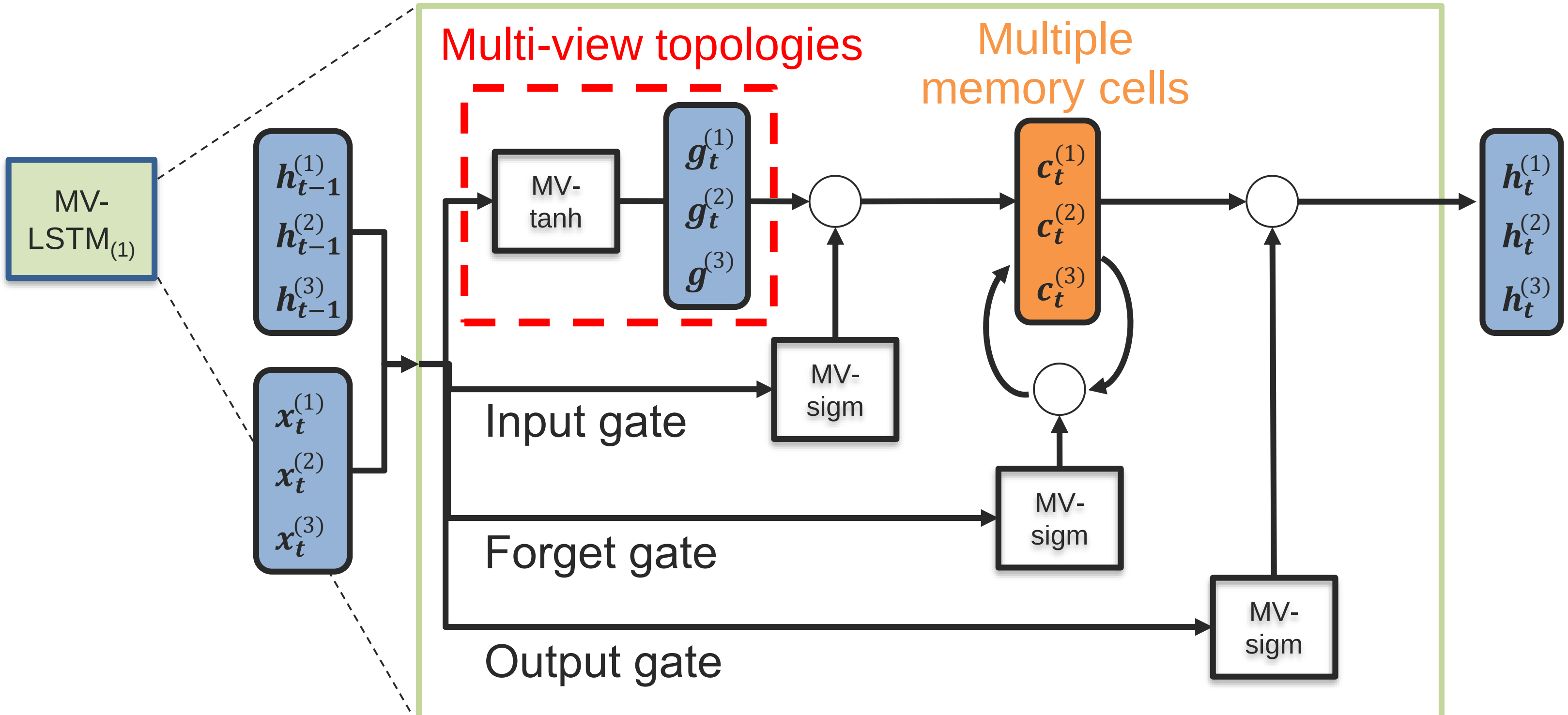


Multi-View Long Short-Term Memory (MV-LSTM)



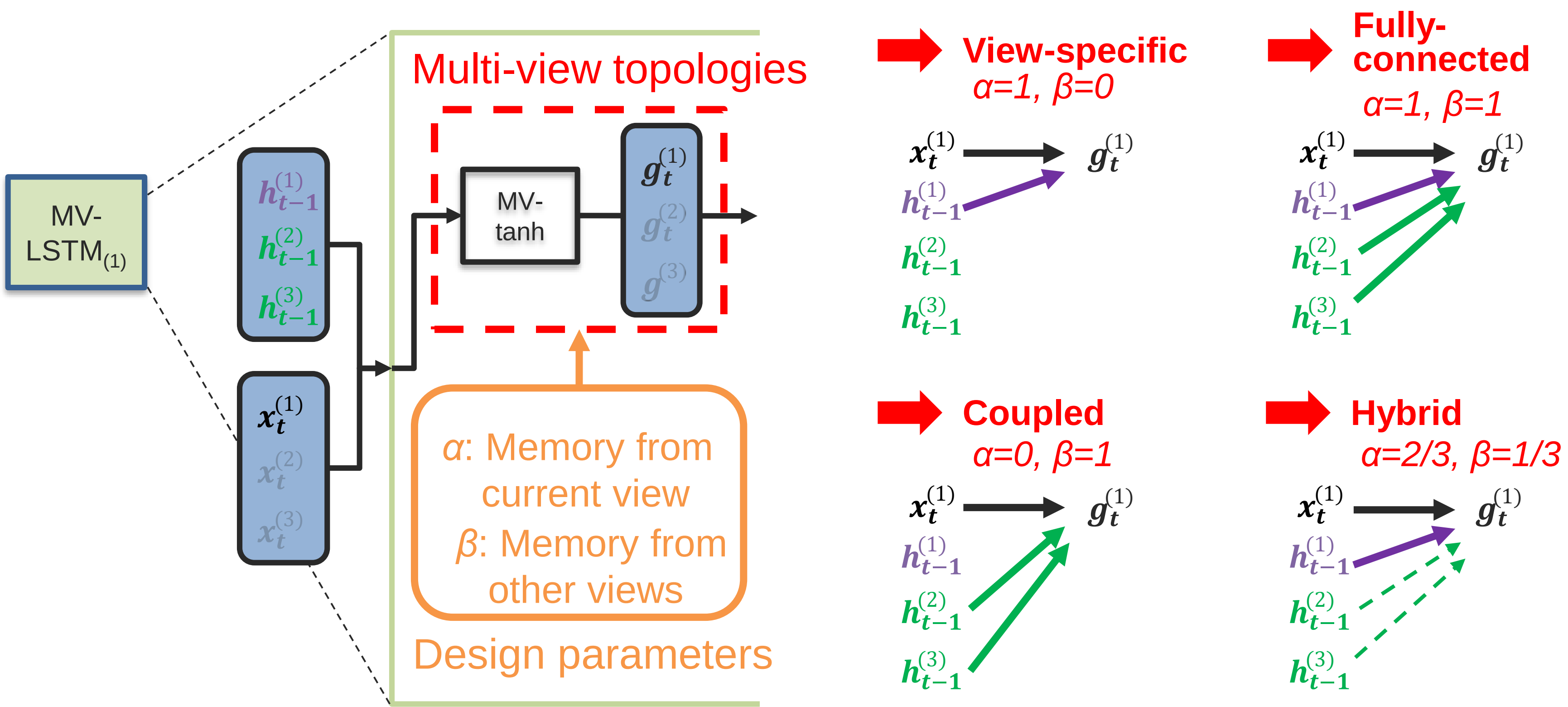
[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

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

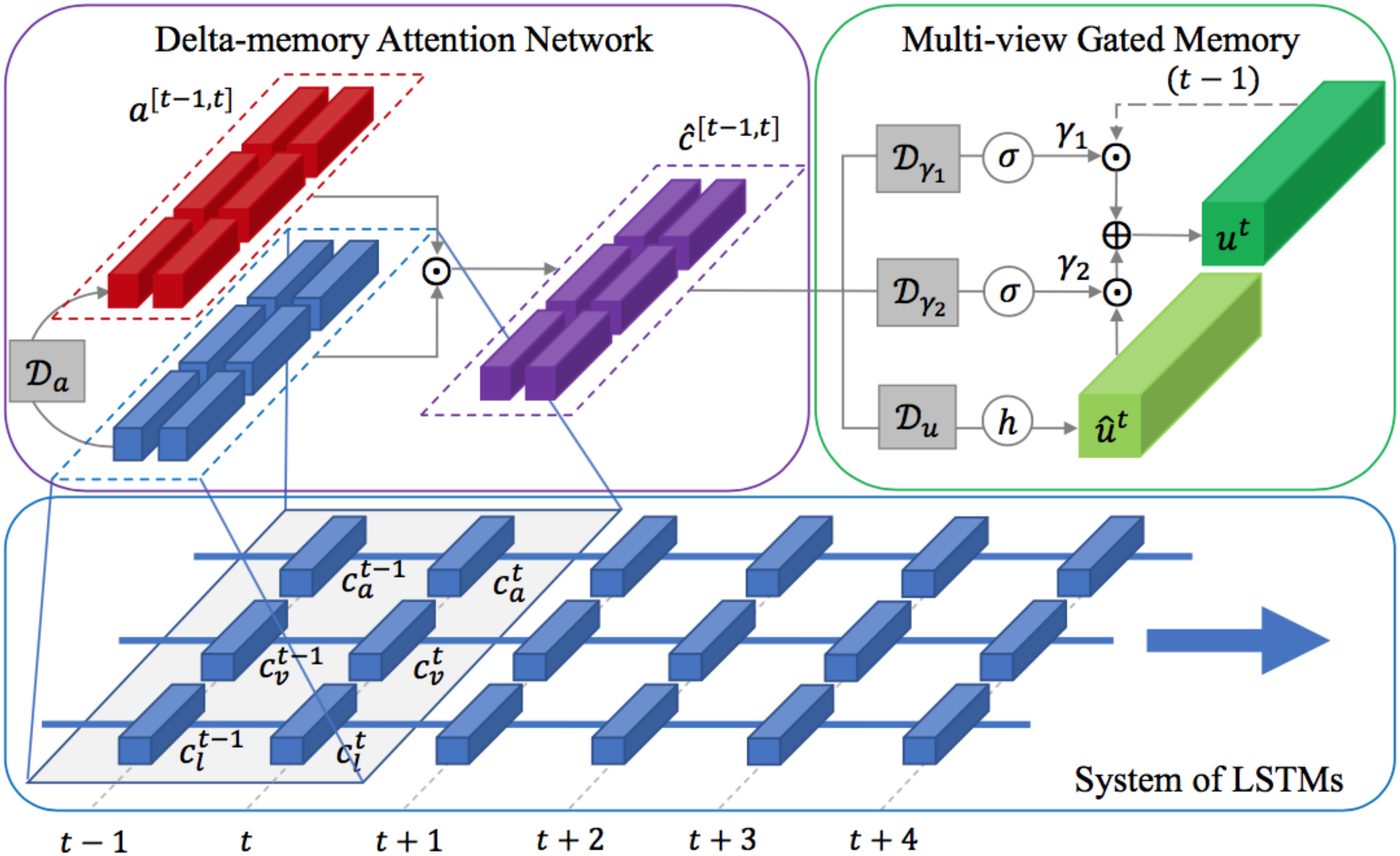
Class labels	Model	Precision	Recall	F1
Easy to engage	LSTM (Early fusion)	0.75	0.81	0.78
	MV-LSTM Full	0.81	0.81	0.81
	MV-LSTM Coupled	0.79	0.81	0.80
	MV-LSTM Hybrid	**0.80**	**0.86**	**0.83**
Difficult to engage	LSTM (Early fusion)	0.63	0.55	0.59
	MV-LSTM Full	0.68	0.68	0.68
	MV-LSTM Coupled	0.67	0.64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