Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

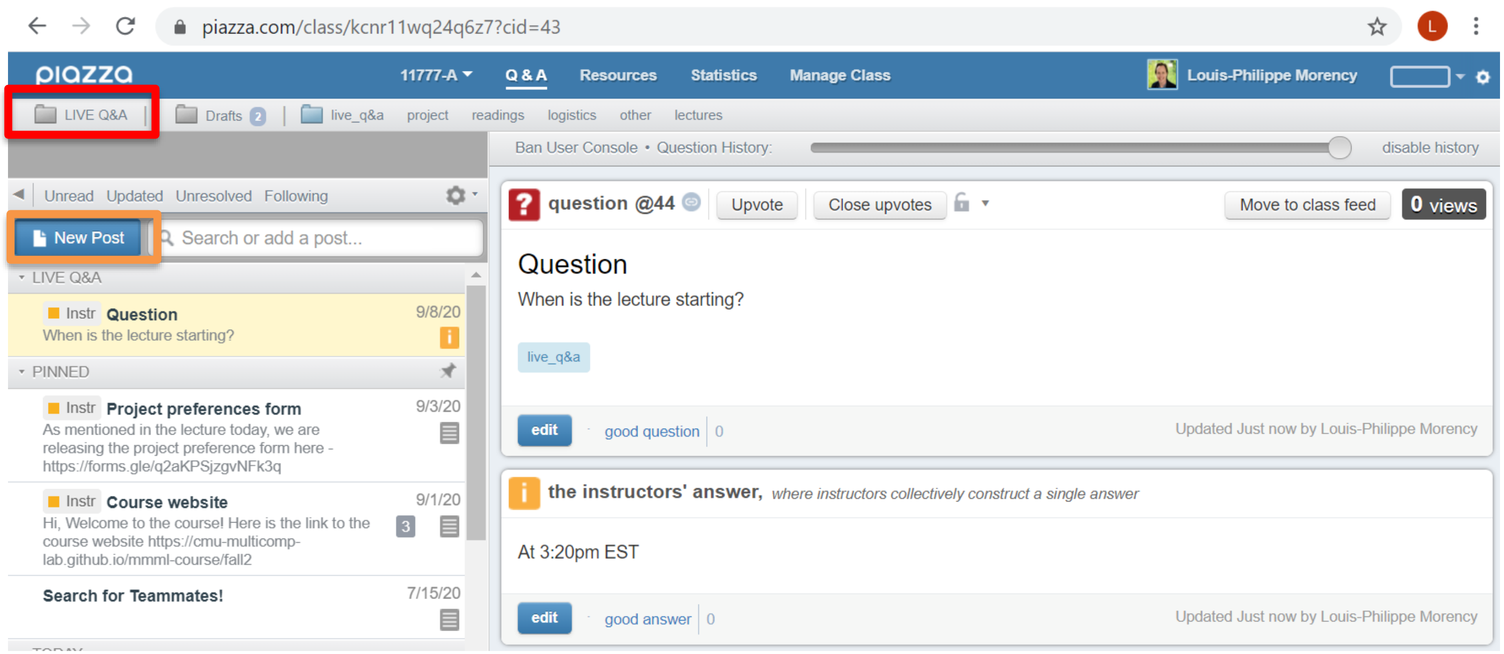
Lecture 4.1: Multimodal Representations

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Administrative Stuff

Piazza Live Q&A – Reminder



Upcoming Schedule

First project assignment:

- Proposal presentations (Friday 10/9)
- First project reports (Sunday 10/11)

Midterm project assignment

- Midterm presentations (Friday 11/12)
- Midterm reports (Sunday 11/14)

Final project assignment

- Final presentations (Friday 12/11)
- Final reports (Sunday 12/13)

Project Proposal Report

Part 1 (updated version of your pre-proposal)

- **Introduction:**
  - Describe and motivate the research problem
  - Define in generic terms the main computational challenges
- **Experimental Setup:**
  - Describe the dataset(s) you are planning to use for this project.
  - Describe the input modalities and annotations available in this dataset.

Project Proposal Report

Part 2

- **Related Work:**
  - Include 12-15 paper citations which give an overview of the prior work
  - Present in more details the 3-4 research papers most related to your work
- **New Research Ideas**
  - Describe your specific challenges and/or research hypotheses
  - Highlight the novel aspect of your proposed research

Project Proposal Report

Part 3

- **Language Modality Exploration:**
  - Explore at least two different computational representations for your language data
  - visualize your language data in relation with your labels
  - Include qualitative examples of successes and failure cases.
- **Visual Modality Exploration:**
  - Explore pre-trained Convolutional Neural Networks (CNNs) on your dataset
  - Load a pre-existing CNN model trained for object recognition (e.g., VGG-Net) and process your test images.
  - Visualize the visual representations (e.g., using t-sne visualization) with overlaid class labels with different colors.

Proposal Presentation

- Presentation should focus on the new research ideas
  - Pre-recorded, 6 minutes maximum (about 5-8 slides)
  - All team members should be involved in the presentation
- Will receive feedback from instructors and other students
  - Peer review process described in next slides
- Submission:
  - Submit your recorded video (MP4) on box.com [LINK]
  - Submit your slides (PDF) on gradescope
- Deadline: Friday 10/9 (on Gradescope)

Peer Feedback

- All videos will be shared on Piazza
- Each video gets a separate post
  - Accessible by all students, TAs and instructor - can share comments, questions and suggestions
- Each student expected to watch at least 6 videos
  - Post feedback for each video (120+ words)
  - Feedback should focus on the new research ideas
- Details about matching will be shared via Piazza
  - Deadline for peer feedback: Friday 10/16

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.1: Multimodal Representations

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

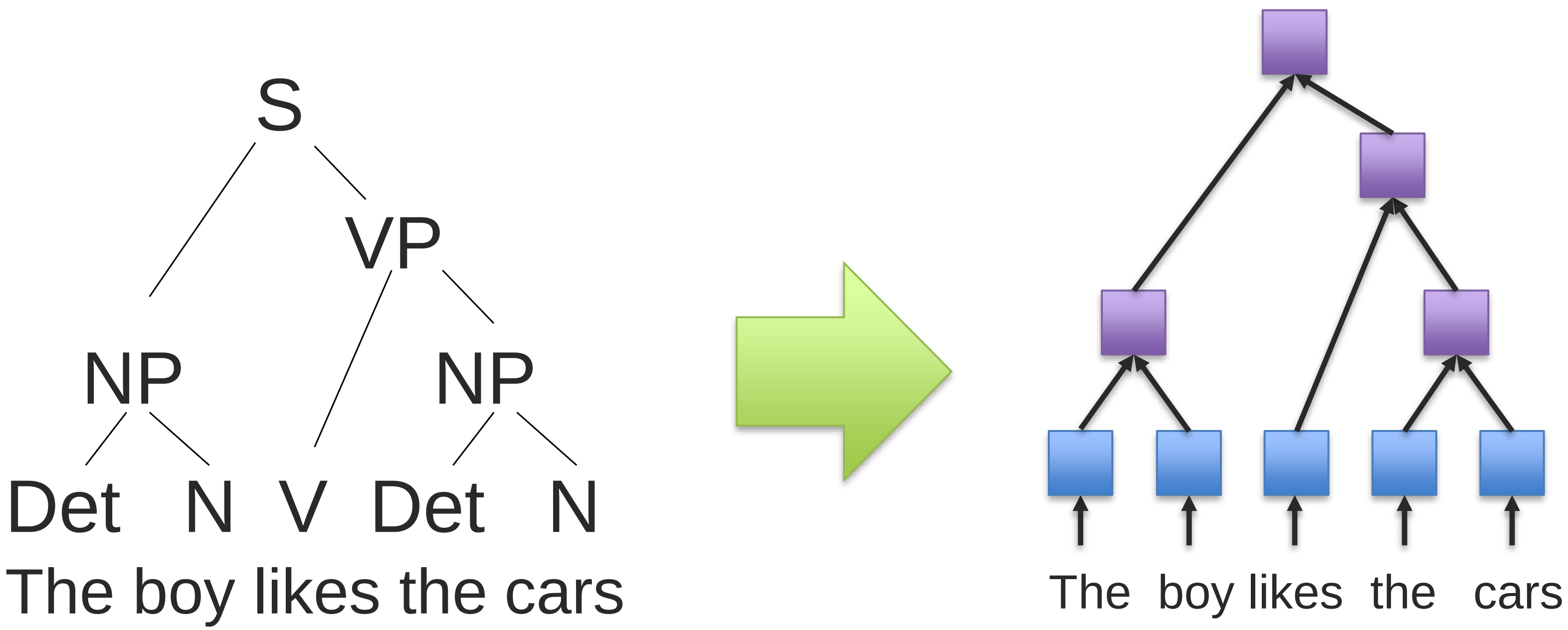
Objectives of today's class

- Unimodal representation
  - Graph-based representations
    - Graph convolution network
- Multi-modal representations
  - Coordinated vs. joint representations
  - Multimodal autoencoders
    - Multimodal Deep Boltzmann Machines
  - Tensor Fusion representation
  - Low-rank fusion representations
  - Multimodal LSTM

Graph Representations

*slides adapted from Leskovec, Representation Learning on Networks. WWW 2018

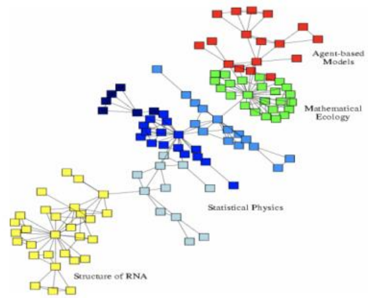
RECAP: Tree-based RNNs (or Recursive Neural Network)



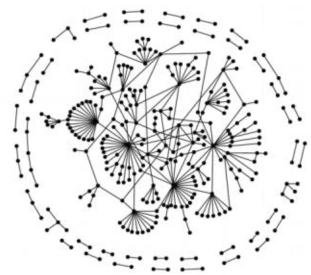
But how to model data with graph-based relations?

Graphs (aka “Networks”)

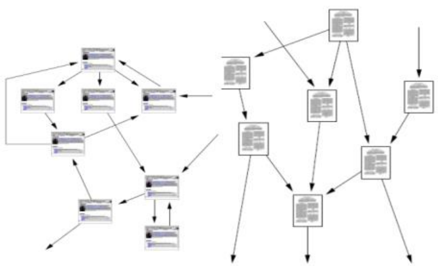
Social networks



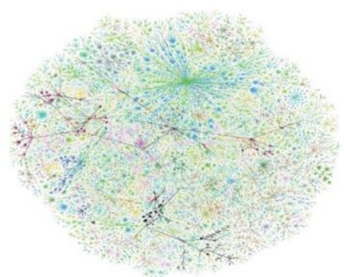
Economic networks



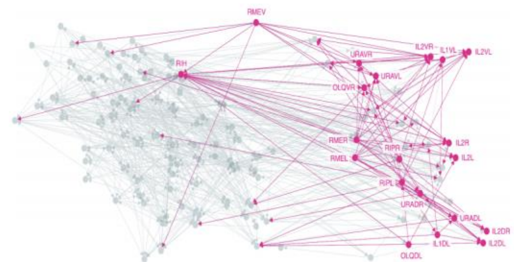
Biomedical networks



Information networks:
Web & citations



Internet

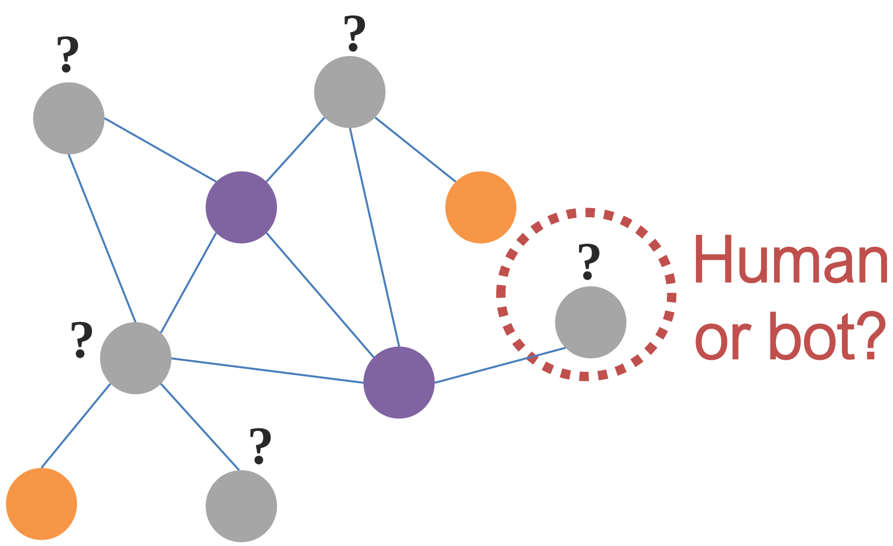


Networks of neurons

Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019

Graphs – Supervised Task

Goal: Learn from labels associated with a subset of nodes (or with all nodes)

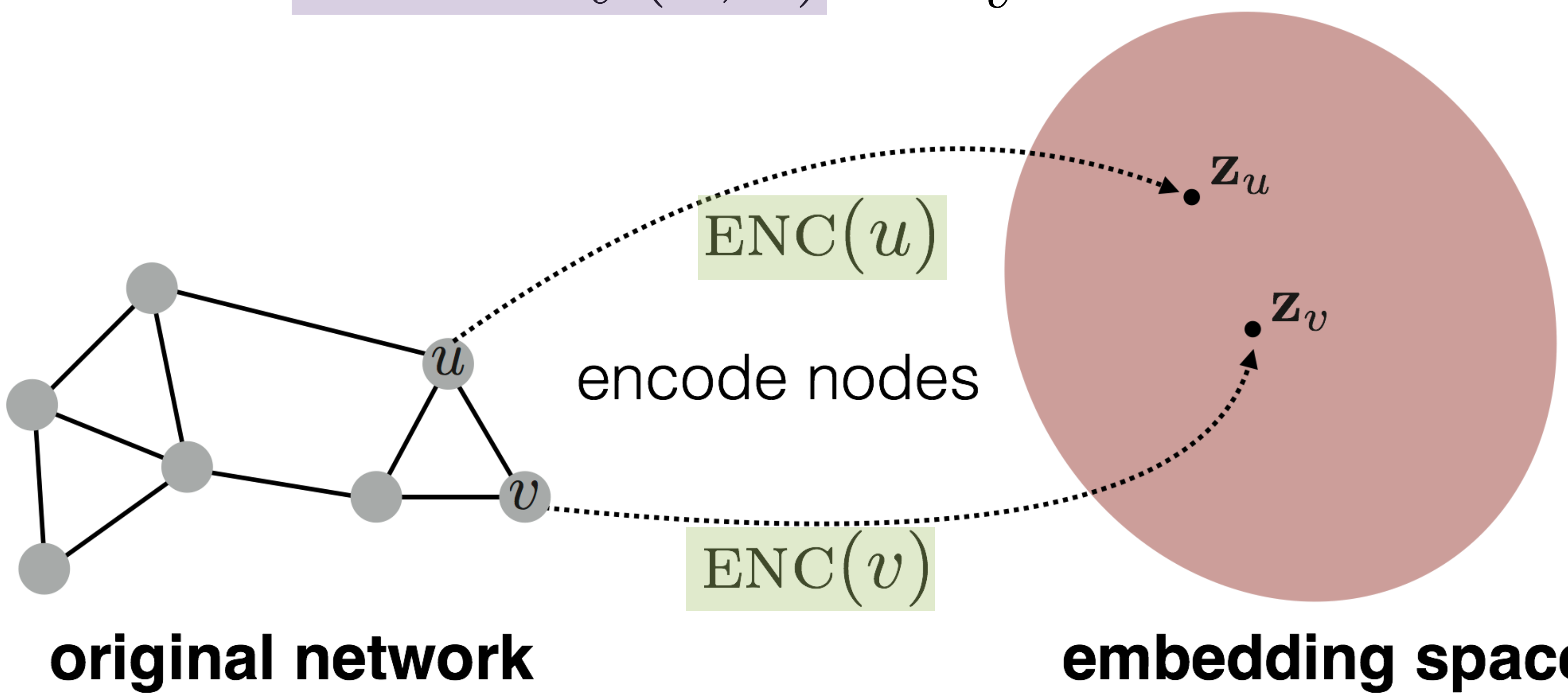


e.g., an online social network

Graphs – Unsupervised Task

Goal: Learn an embedding space where

$$\text{similarity}(u, v) \approx \mathbf{z}_v^\top \mathbf{z}_u$$



Graph Neural Nets

Assume we have a graph **G:**

V is the set of vertices

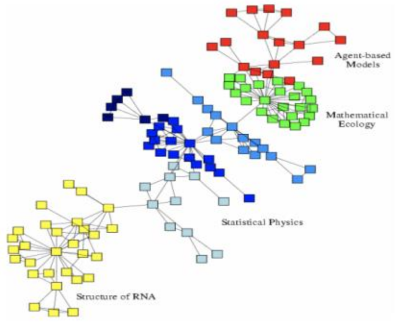
A is the binary adjacency matrix

X is a matrix of node features:

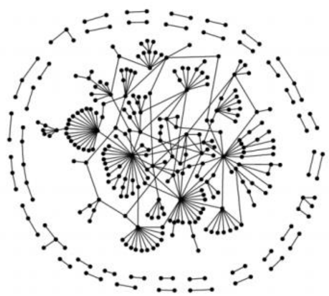
- Categorical attributes, text, image data
  e.g. profile information in a social network
- …

Y is a vector of node labels (optional)

Social networks



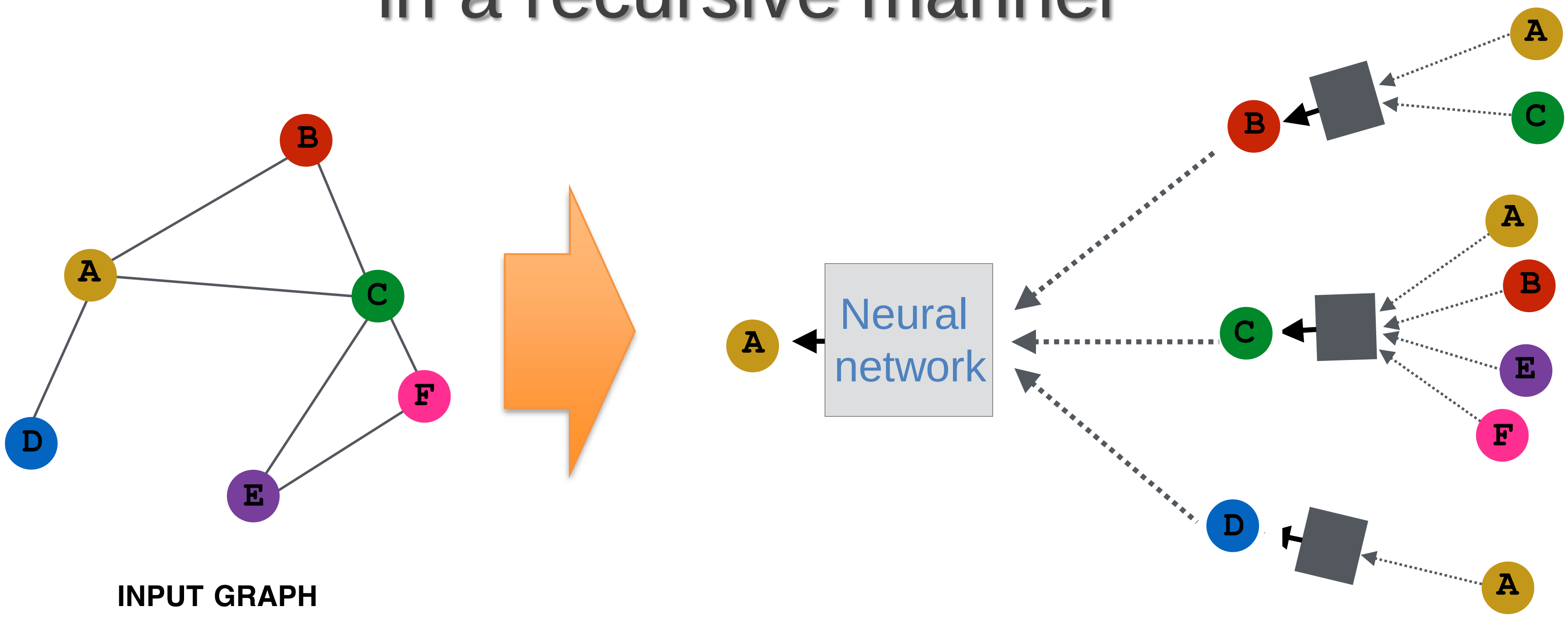
Economic networks



Biomedical networks

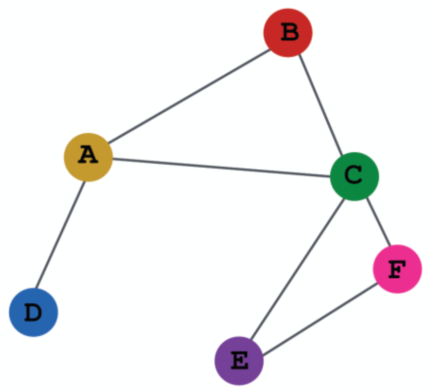
Graph Neural Nets

Key idea: Generate node embeddings based on local neighborhoods in a recursive manner

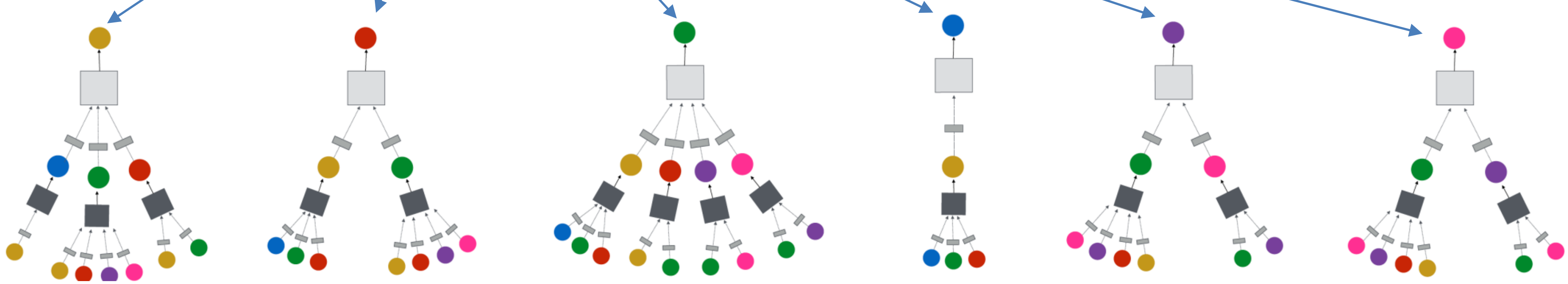


Graph Neural Nets

Every node defines a unique computation graph!



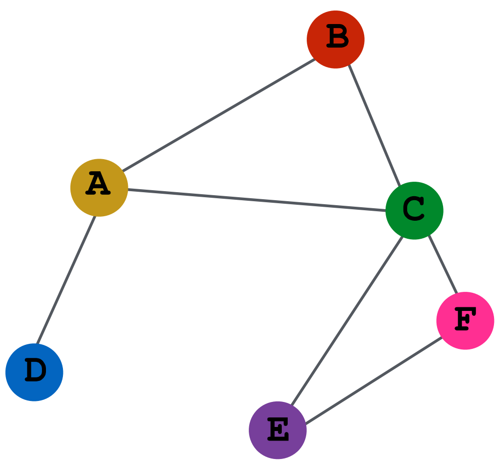
INPUT GRAPH



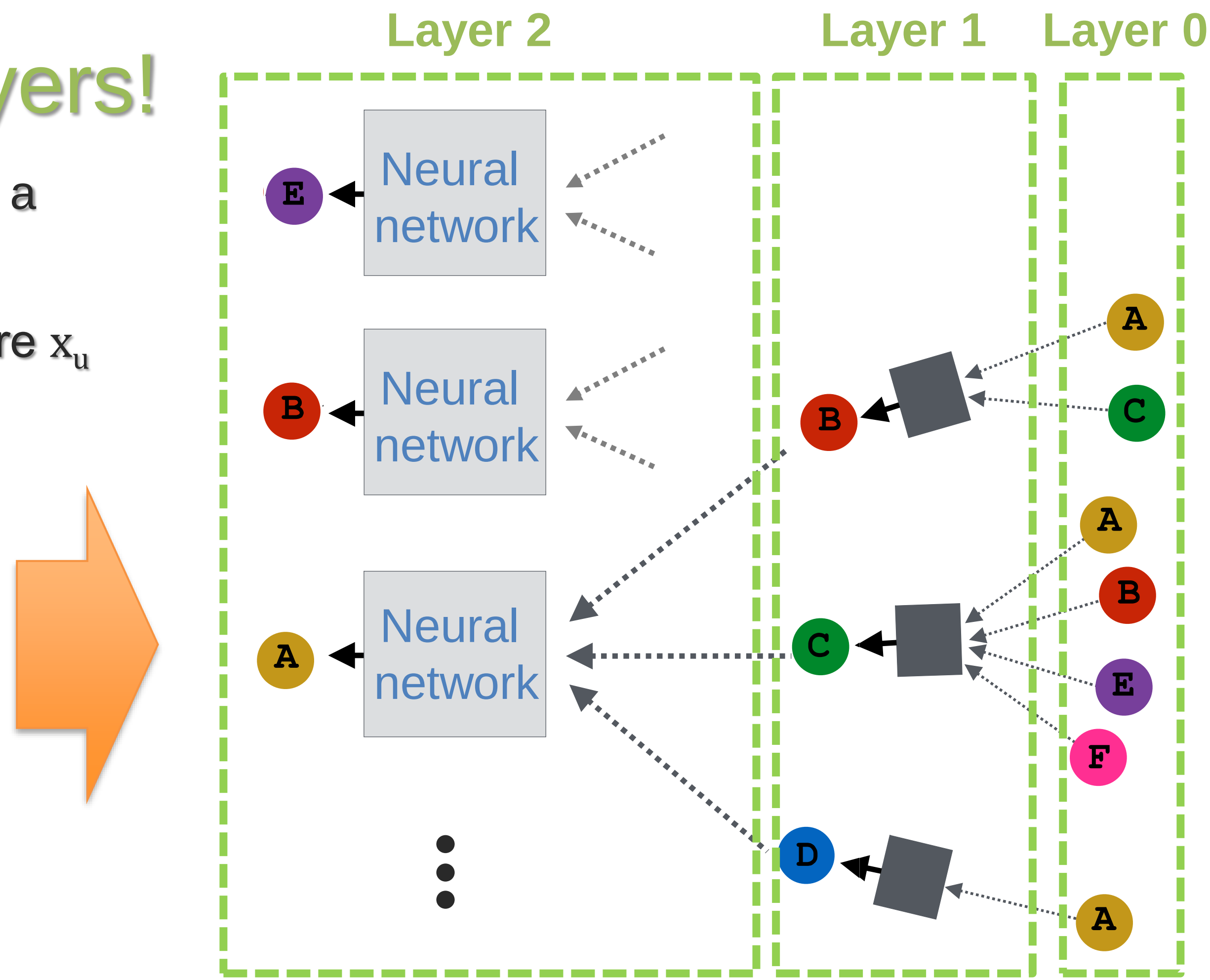
Graph Neural Nets

And multiple layers!

- Shared parameters within a specific layer
- “layer-0” is the input feature x_u



INPUT GRAPH

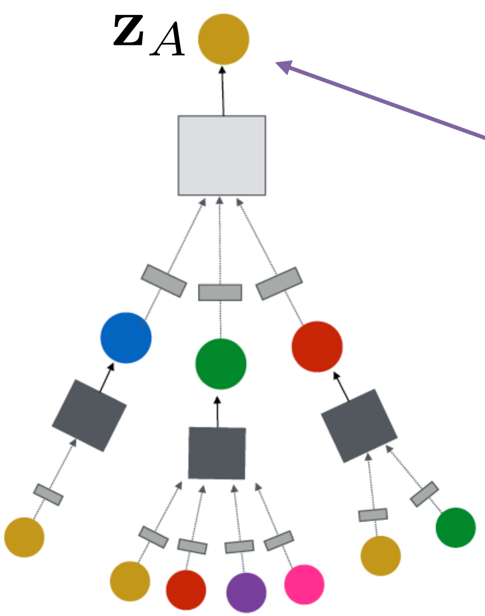


Graph Neural Nets – Supervised Training

Human or bot?

classification weights

$$\mathcal{L} = \sum_{v \in V} \boxed{y_v} \log(\sigma(\boxed{\mathbf{z}_v^\top}\boxed{\theta})) + (1 - \boxed{y_v}) \log(1 - \sigma(\boxed{\mathbf{z}_v^\top}\boxed{\theta}))$$

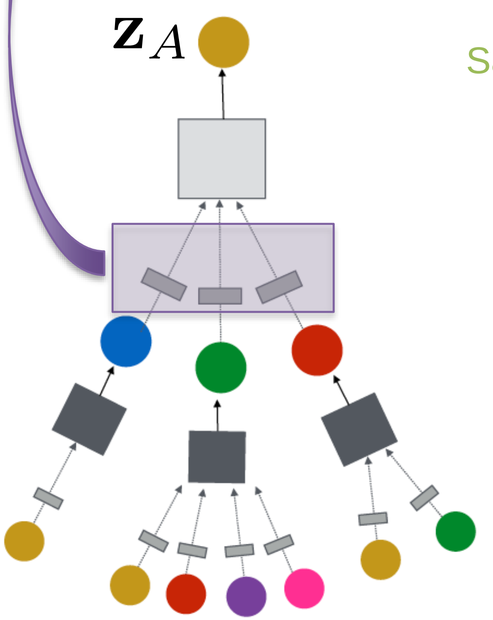


output node embedding

node class label

Graph Neural Nets – Neighborhood Aggregation

How to aggregate multiple neighbors ?



Average pooling (Scarselli et al., 2005)

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1}\right)$$

Different weights for neighbors and self

Graph Convolution Network (Kipf et al., 2017)

Same weights

$$\mathbf{h}_v^k = \sigma\left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}}\right)$$

Different normalization

- It can be efficiently implemented

Gated Graph Neural Network (Li et al., 2016)

Same weights across layers

$$\mathbf{h}_v^k = \mathrm{GRU}(\mathbf{h}_v^{k-1}, \mathbf{m}_v^k) \quad \text{where} \quad \mathbf{m}_v^k = \mathbf{W} \sum_{u \in N(v)} \mathbf{h}_u^{k-1}$$

- It can handle deeper networks

Graph Neural Nets – References

Graph Conv Nets

Kipf et al., 2017. Semi-supervised Classification with Graph Convolutional Networks. ICLR.

Gated Graph Nets

Li et al., 2016. Gated Graph Sequence Neural Networks. ICLR.

Subgraph embeddings

Duvenaud et al. 2016. Convolutional Networks on Graphs for Learning Molecular Fingerprints. ICML.

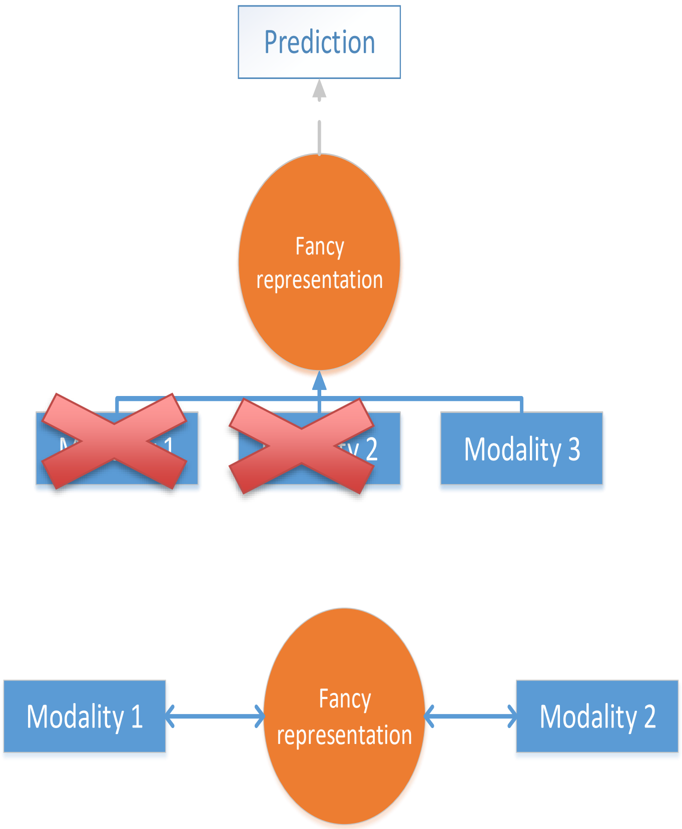
Li et al. 2016. Gated Graph Sequence Neural Networks. ICLR.

Multimodal representations

Multimodal representations

What do we want from multi-modal representation?

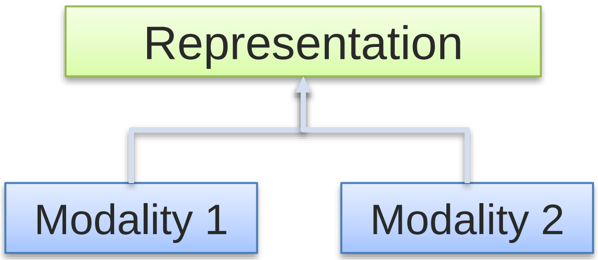
- Similarity in that space implies similarity in corresponding *concepts*
- Useful for various discriminative tasks – retrieval, mapping, fusion etc.
- Possible to obtain in absence of one or more modalities
- Fill in missing modalities given others (map between modalities)



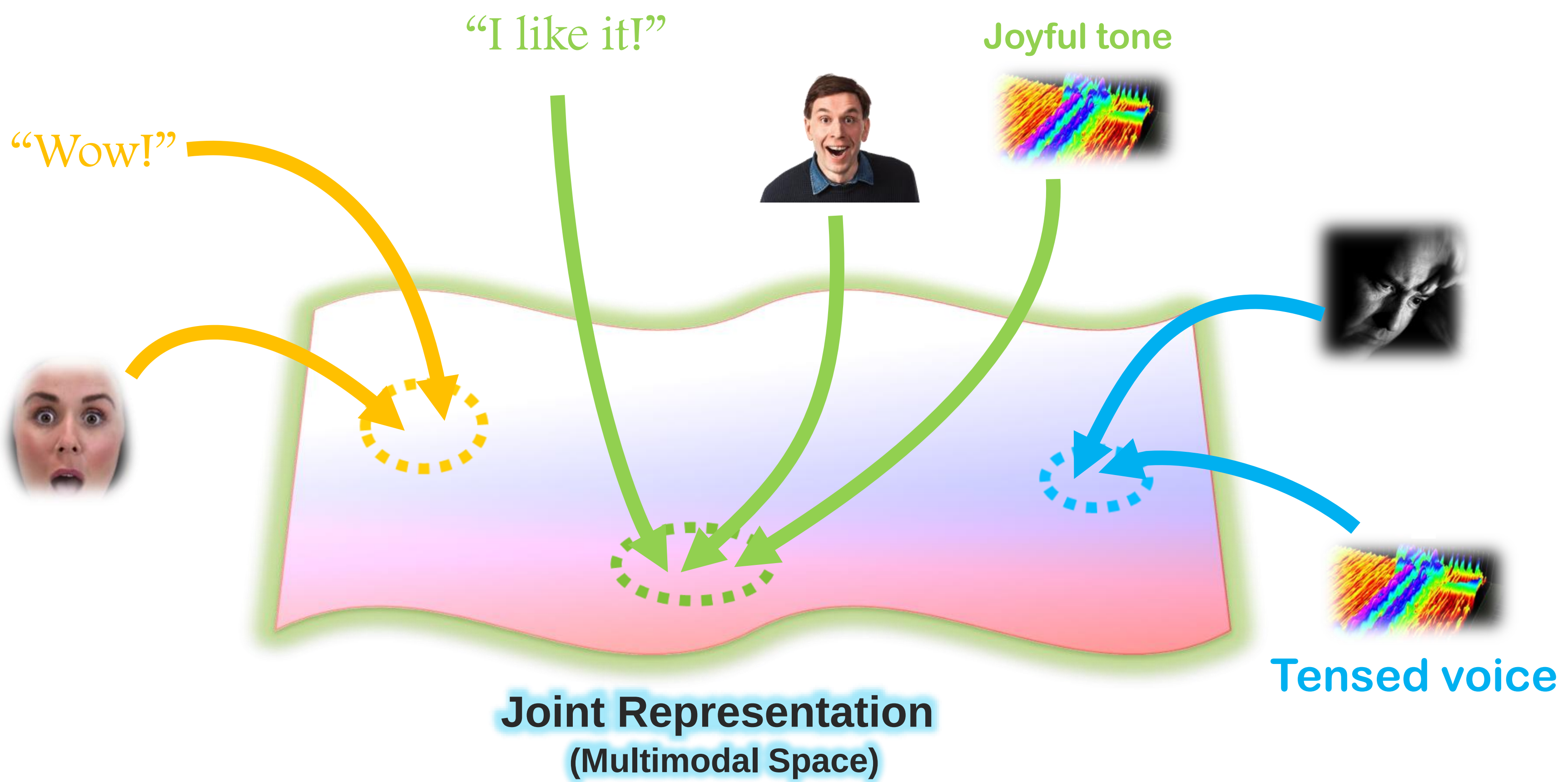
Core Challenge: Multimodal Representation

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

A Joint representations:



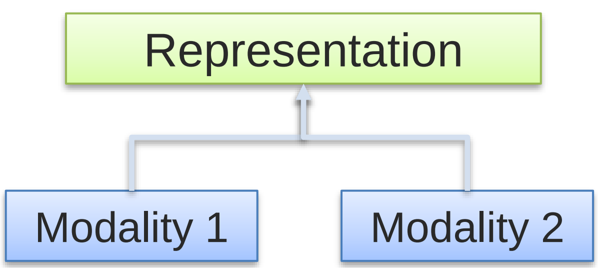
Joint Multimodal Representation



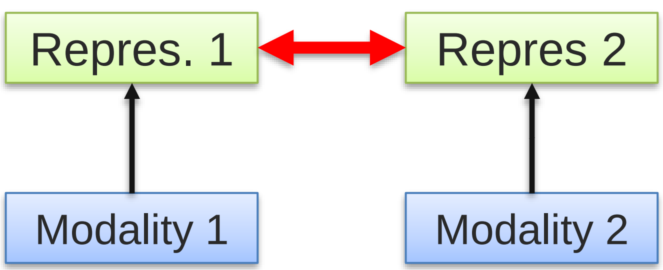
Core Challenge 1: Representation

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

A Joint representations:



B Coordinated representations:



Unsupervised Joint representations

Unsupervised learning

Unlabeled data $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ …

… with no labels $Y = \{y_1, y_2, \ldots, y_n\}$

Why would we want to tackle such a task?

1. Extracting interesting information from data
   - Clustering
   - Discovering interesting trends
   - Data compression
2. Learn better representations

Unsupervised representation learning

Force our representations to better model input distribution

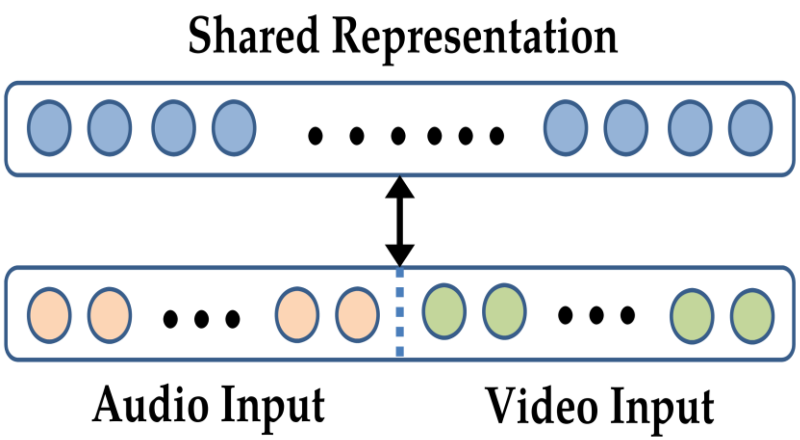
- Not just extracting features for classification
- Asking the model to be good at representing the data and not overfitting to a particular task
- Potentially allowing for better generalizability

Use as initialization for a supervised task, especially when we have a lot of unlabeled data and much less labeled examples

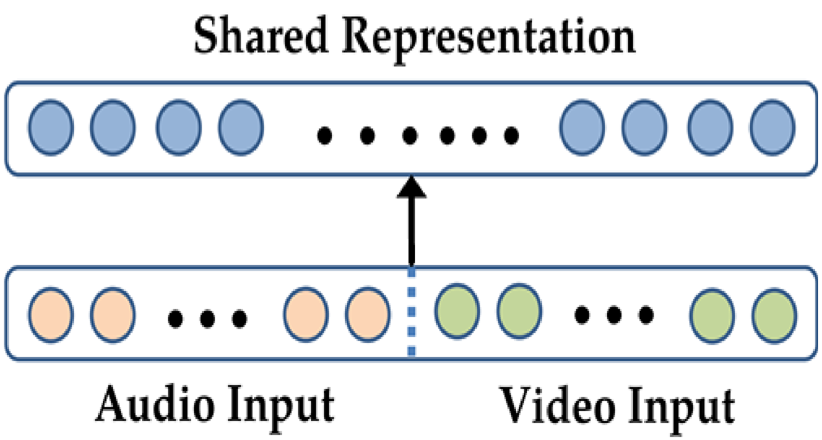
Shallow multimodal representations

Want deep multimodal representations

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



Shallow RBM



Shallow Autoencoder

Autoencoders

What does auto mean?

- Greek for self – self encoding
- Feed forward network intended to reproduce the input

Two parts encoder/decoder

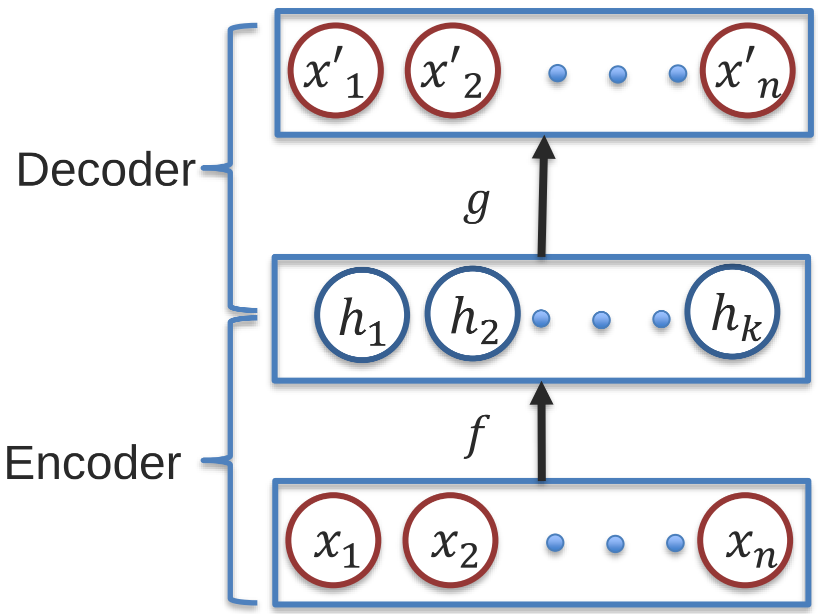
$$\boldsymbol{x}' = \boldsymbol{f}(\boldsymbol{g}(\boldsymbol{x}))$$ **: score function**

$$f = \sigma(W\boldsymbol{x})$$: encoder

$$g = \sigma(W^*\boldsymbol{h})$$: decoder

Often, we use *tied weights* to force the sharing of weights in encoder/decoder

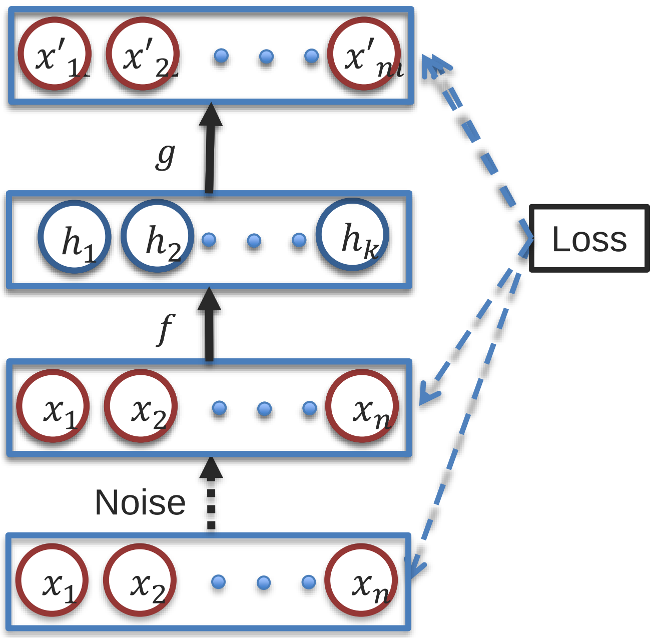
$$W^* = W^T$$



Autoencoder – Loss Function

Loss function compares the original input to the generated output

e.g., Eucledian loss: $L = \frac{1}{2}\sum_k (x_k - x'_k)^2$



But how to make it robust to noise?

Solution: Denoising autoencoder

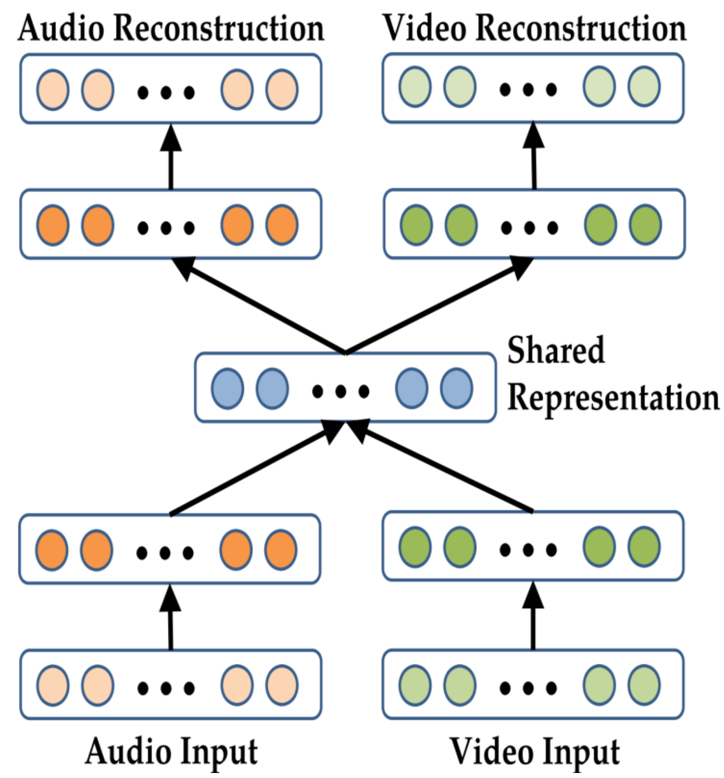
- It adds noise to input $\boldsymbol{x}$ but learn to reconstruct original

It leads to a more robust representation and prevents copying

Deep Multimodal autoencoders

Bimodal auto-encoder: a deep representation learning approach

- Used for Audio-visual speech recognition



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

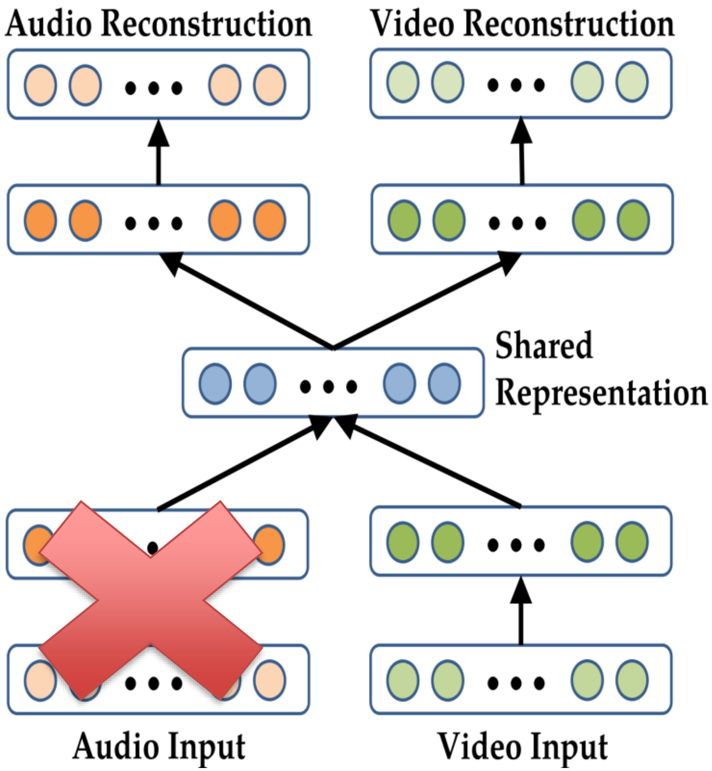
Deep Multimodal autoencoders - training

Individual modalities can be pre-trained

- Denoising Autoencoders

To train the model to reconstruct the other modality

- Use both
- Remove audio



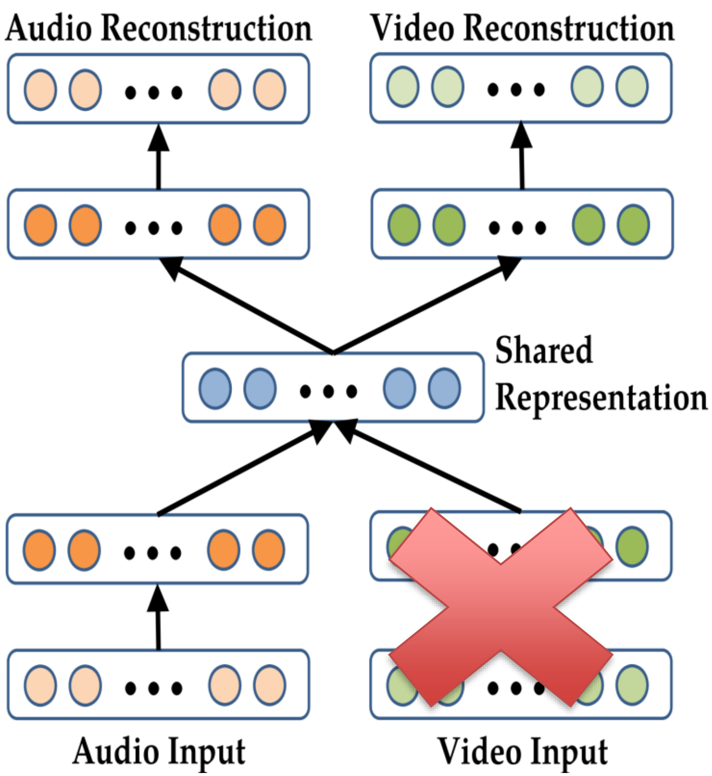
Deep Multimodal autoencoders - training

Individual modalities can be pretrained

- RBMs
- Denoising Autoencoders

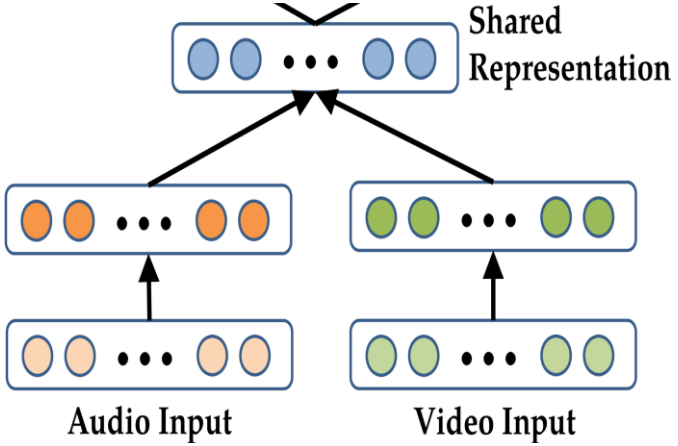
To train the model to reconstruct the other modality

- Use both
- Remove audio
- Remove video



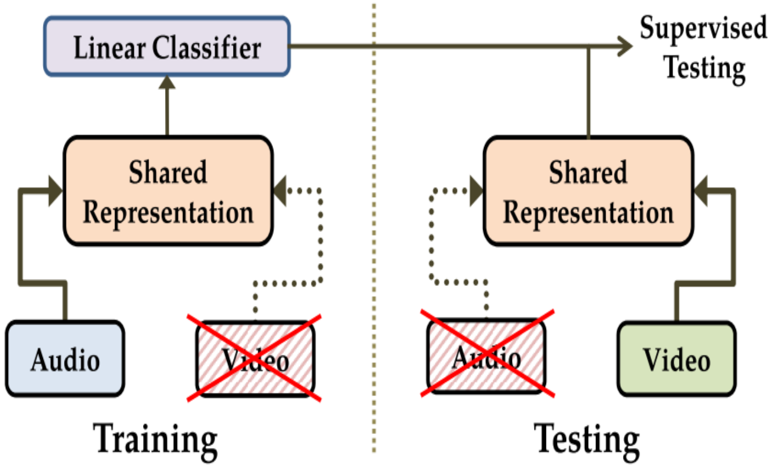
Deep Multimodal autoencoders

It can now discard the decoder and use it for the AVSR task



Interesting experiment:

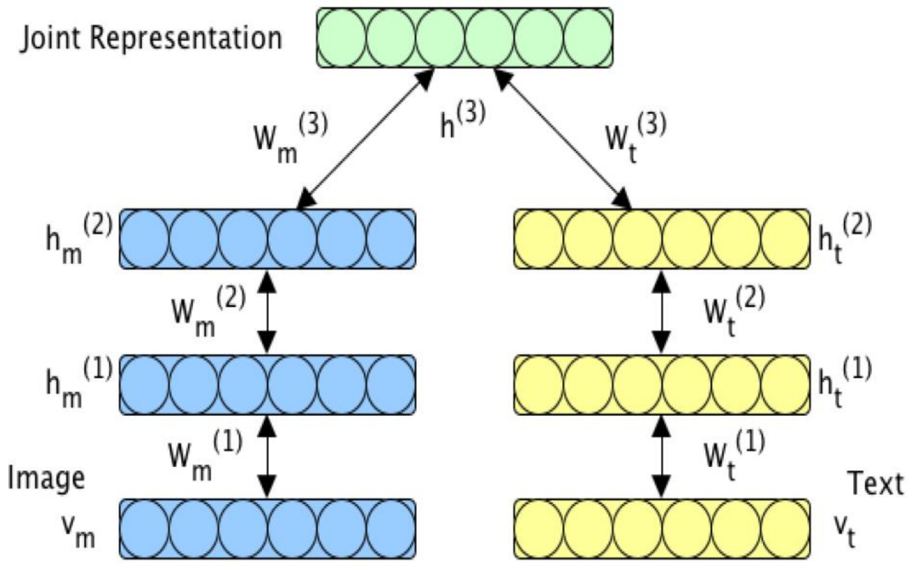
- "Hearing to see"



Deep Multimodal Boltzmann machines

Generative model

- 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]

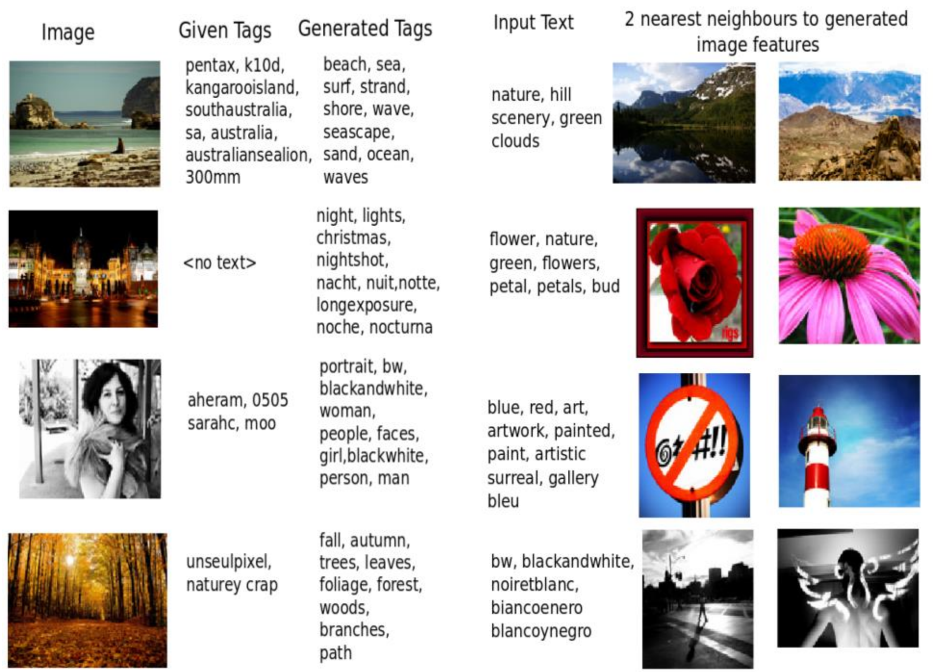
We will discuss Boltzmann machines in more details during lecture 8.1

Deep Multimodal Boltzmann machines

Pre-training on unlabeled data helps

Can use generative models

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)	$\mathbf{0.585} \pm 0.004$	$\mathbf{0.836} \pm 0.004$



Code is available:

http://www.cs.toronto.edu/~nitish/multimodal/

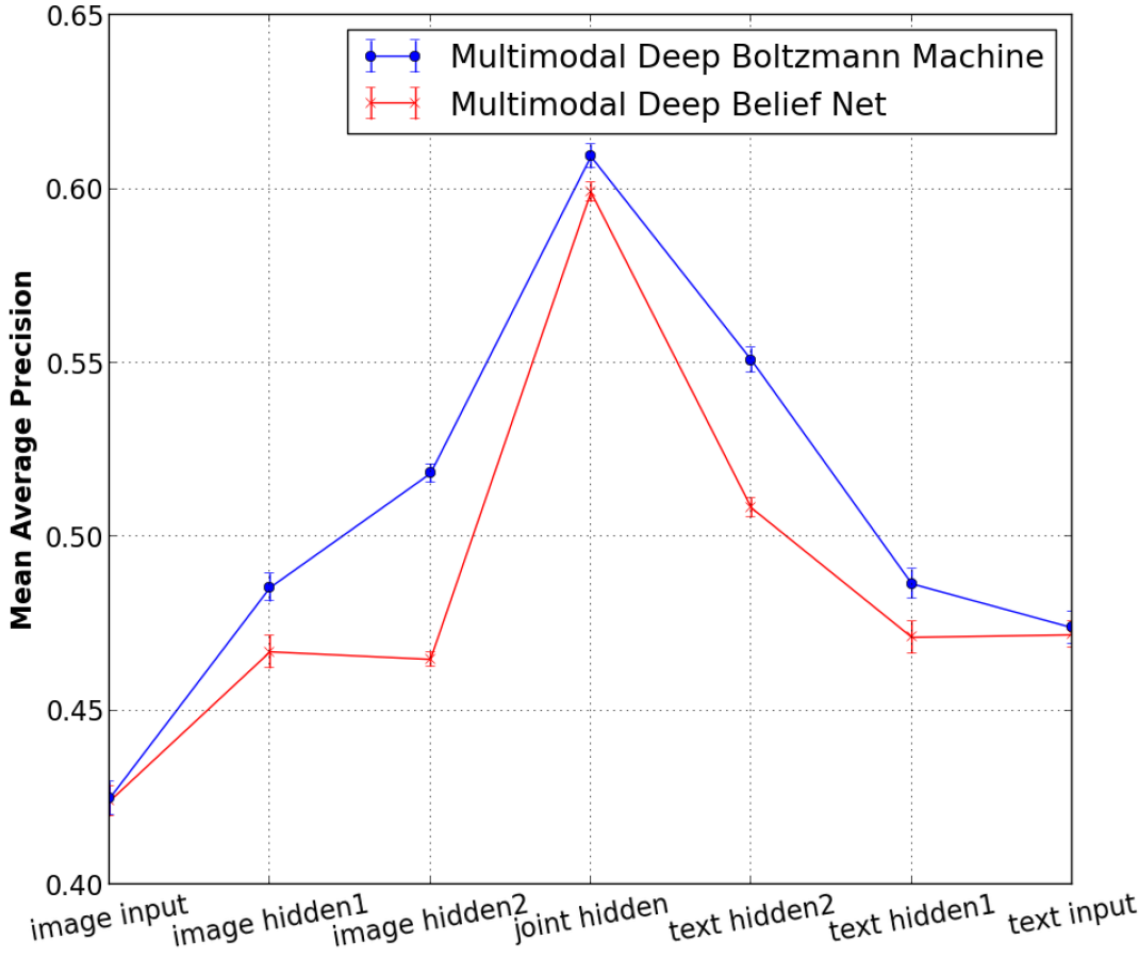
Deep Multimodal Boltzmann Machines

Text information can help visual predictions!

- Image retrieval task on MIR Flickr dataset

Model	MAP	Prec@50
Image LDA (Huiskes et al., 2010)	0.315	-
Image SVM (Huiskes et al., 2010)	0.375	-
Image DBN	0.463 ± 0.004	0.801 ± 0.005
Image DBM	0.469 ± 0.005	0.803 ± 0.005
Multimodal DBM (generated text)	$\mathbf{0.531 \pm 0.005}$	$\mathbf{0.832 \pm 0.004}$

Analyzing Intermediate Representations



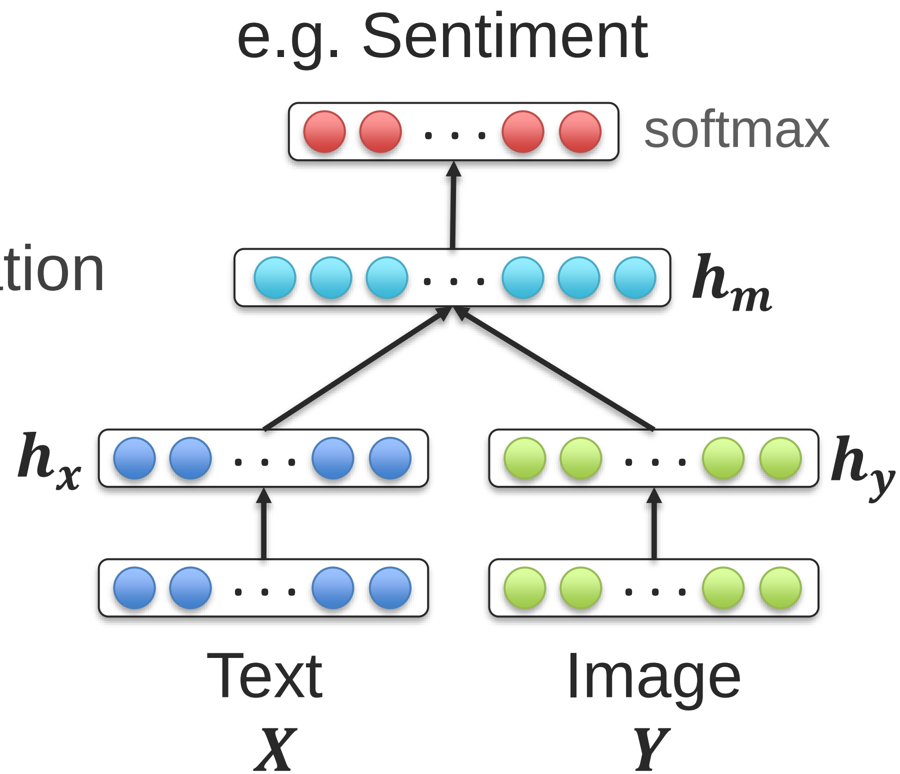
Supervised
Joint representations

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?



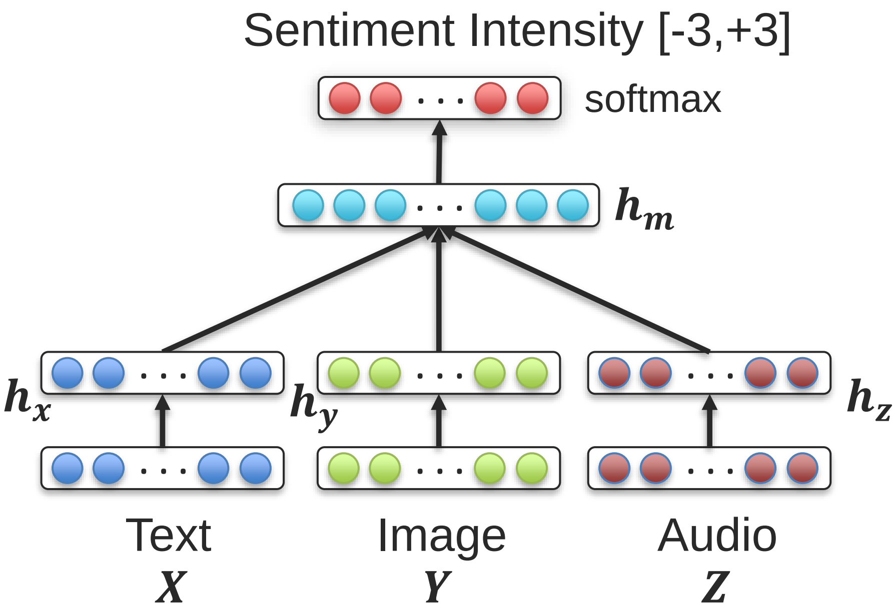
Multimodal Sentiment Analysis

MOSI dataset (Zadeh et al, 2016)

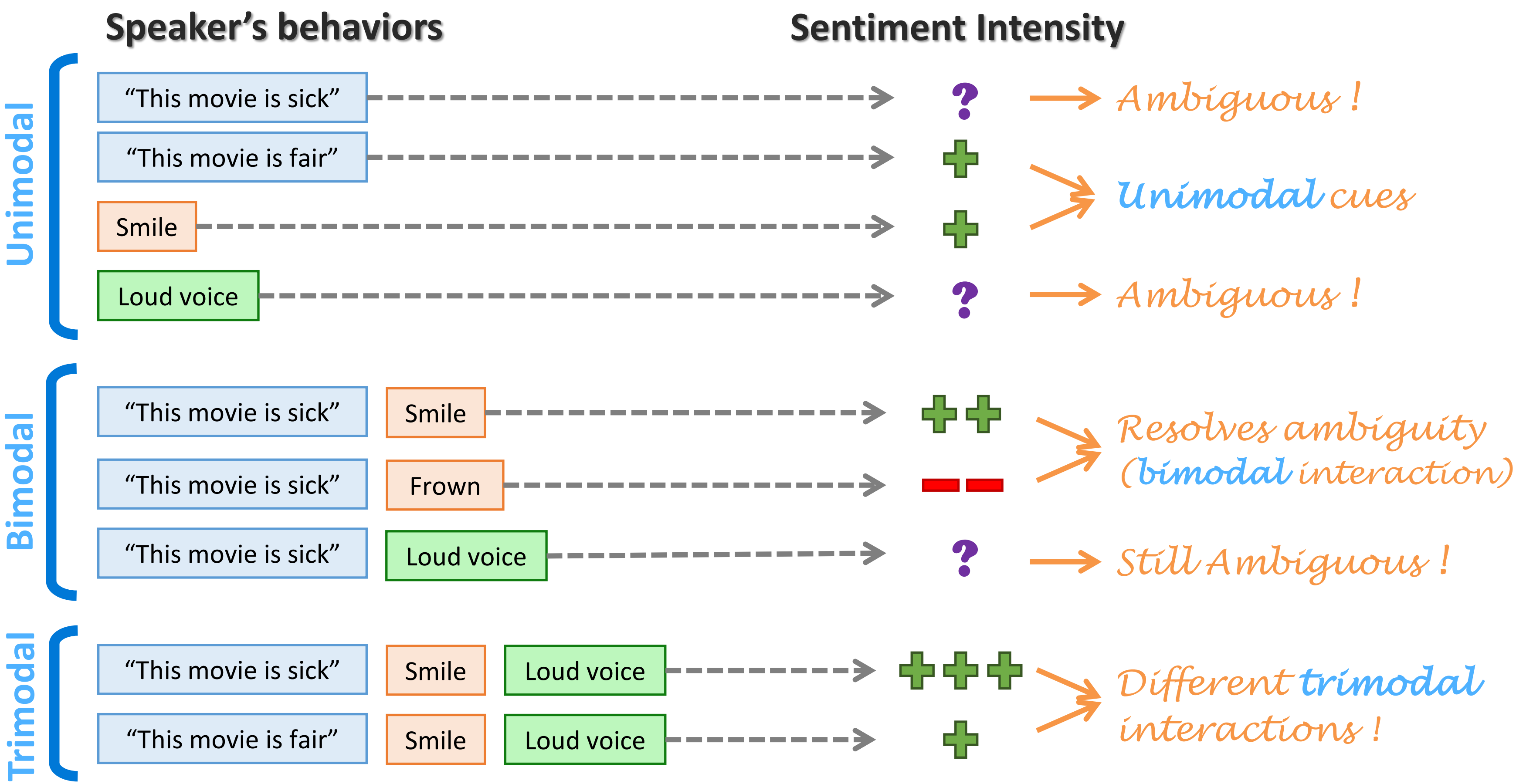
- 2199 subjective video segments
- Sentiment intensity annotations
- 3 modalities: text, video, audio

Multimodal joint representation:

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



Unimodal, Bimodal and Trimodal Interactions

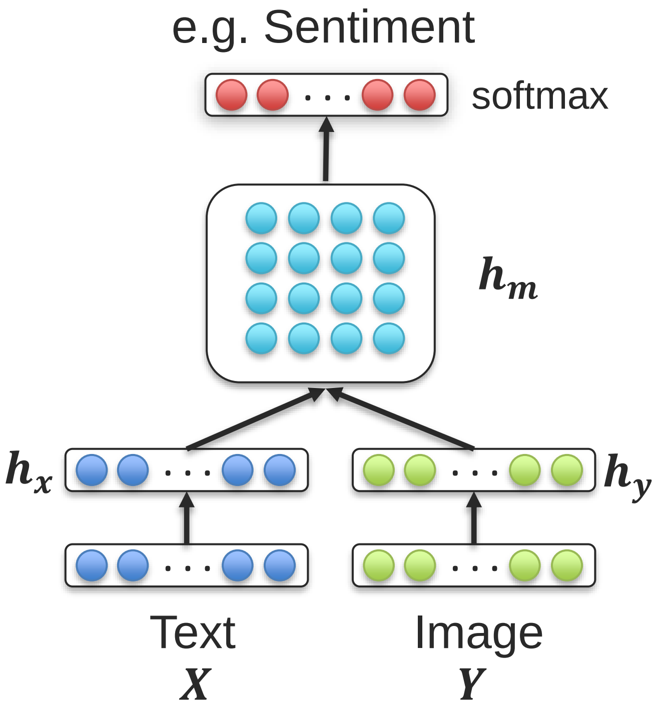


Bilinear Pooling

Models bimodal interactions:

$$\boldsymbol{h_m} = \boldsymbol{h_x} \otimes \boldsymbol{h_y} = \boldsymbol{h_x} \otimes \boldsymbol{h_y}$$

[Tenenbaum and Freeman, **2000**]



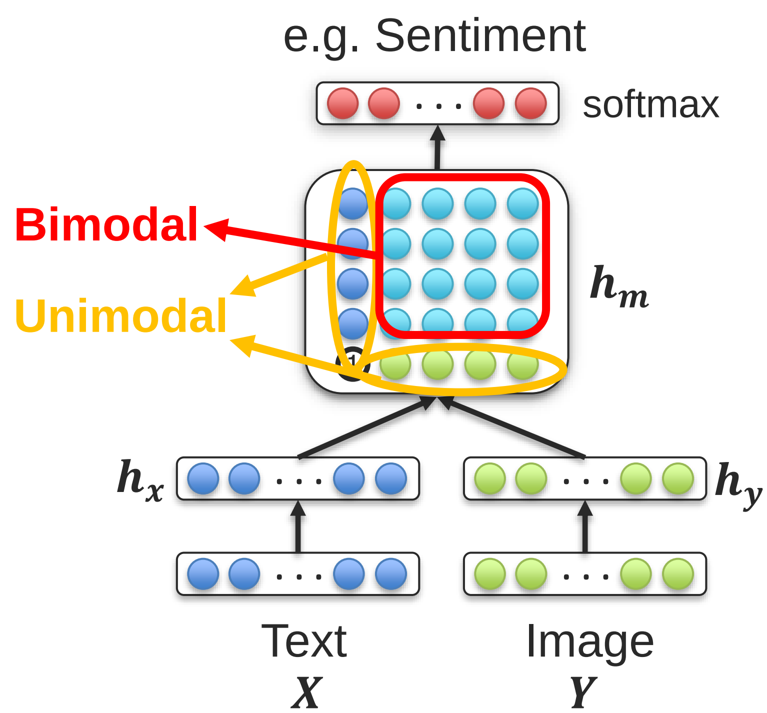
Multimodal Tensor Fusion Network (TFN)

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 !

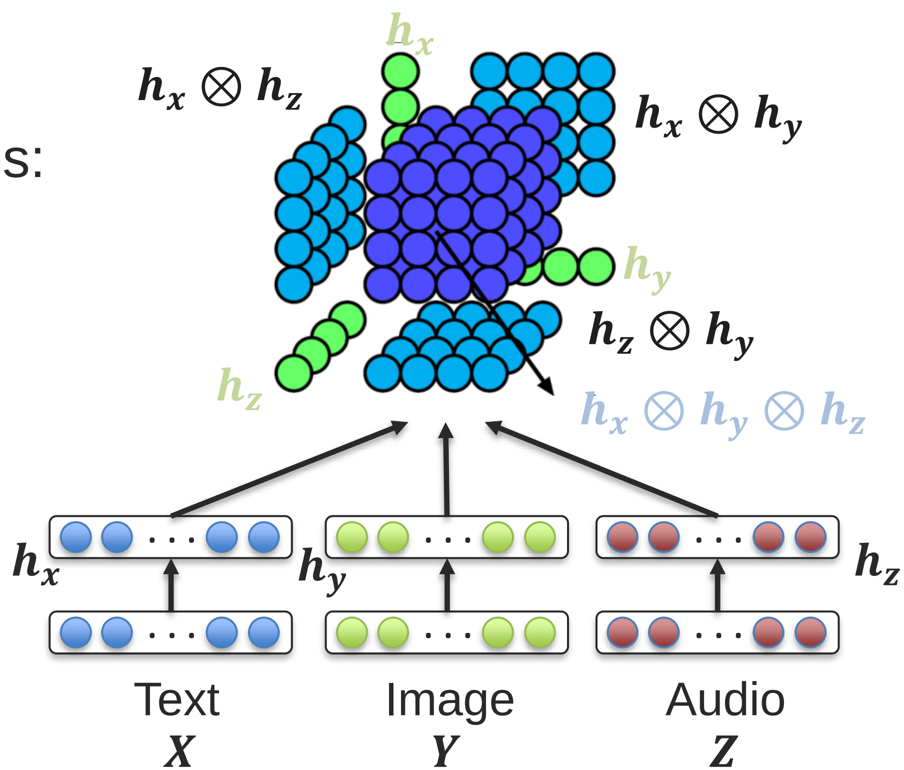
[Zadeh, Jones and Morency, **EMNLP 2017**]



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**]

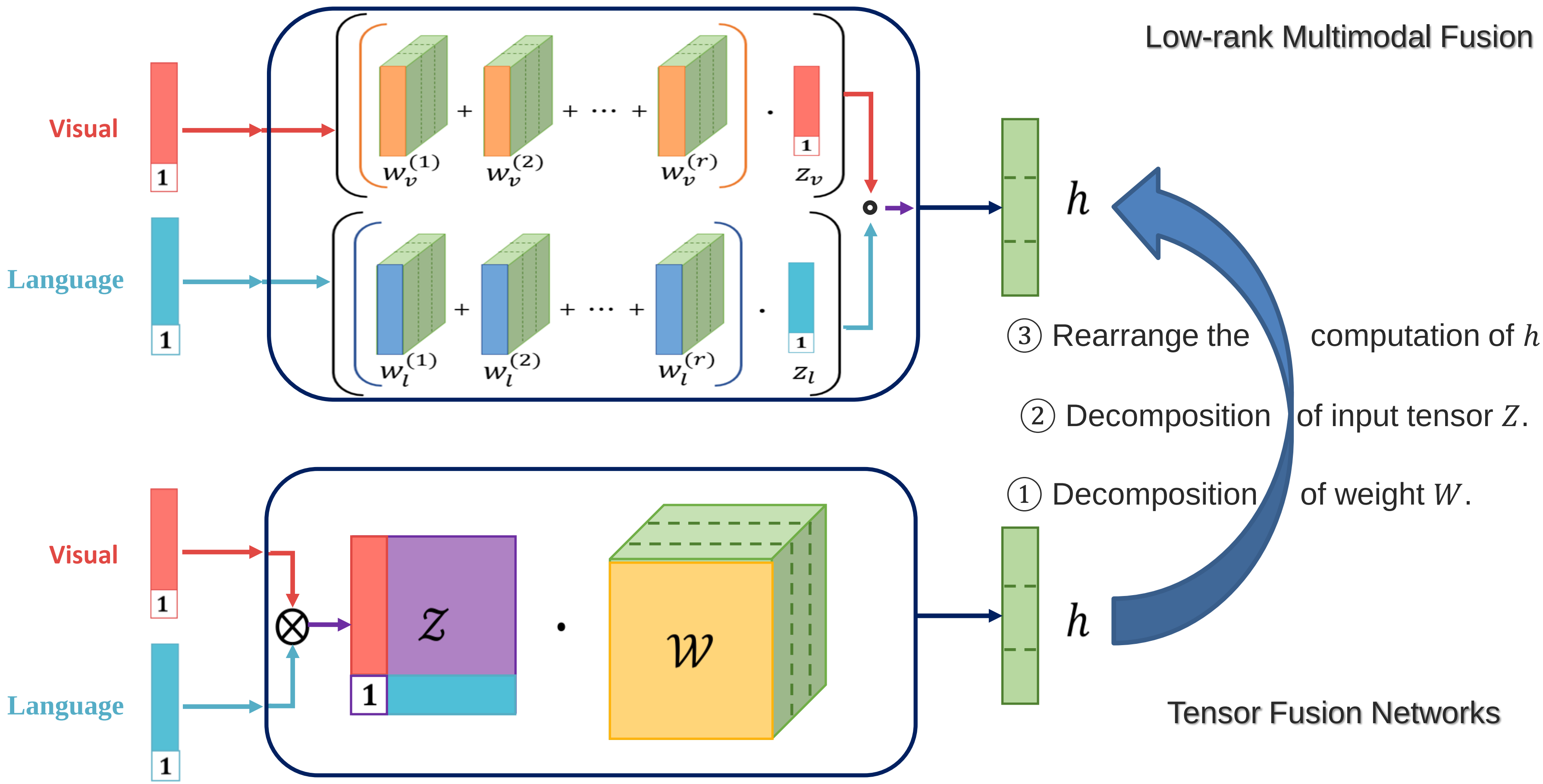
Experimental Results – MOSI Dataset

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

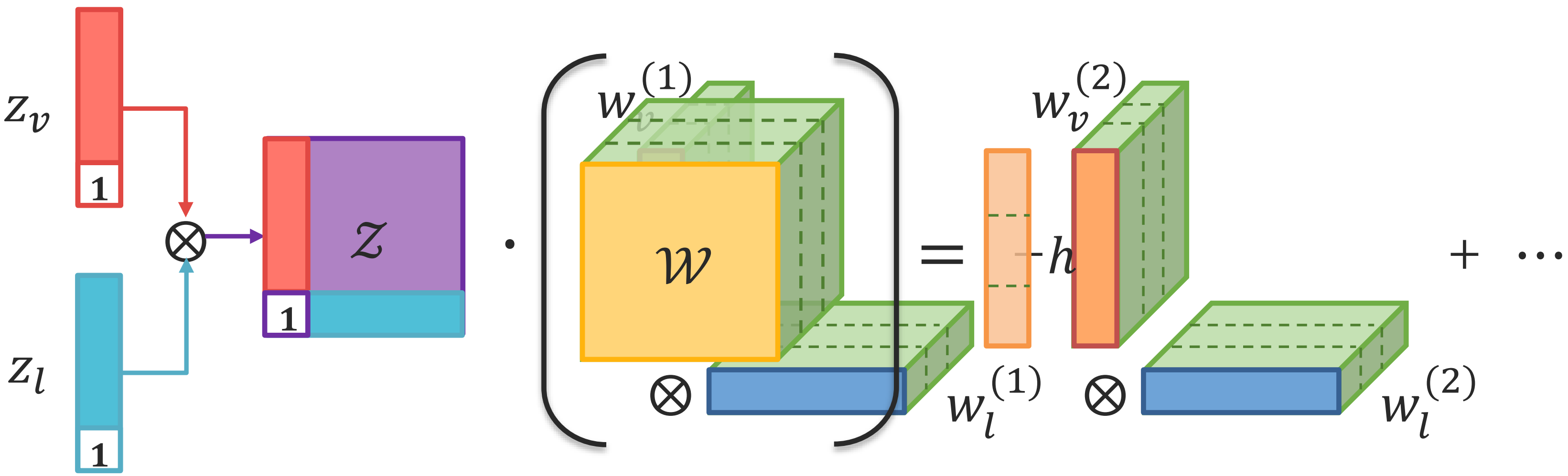
Improvement over State-Of-The-Art

Baseline	Binary		5-class	Regression	
	Acc(%)	F1	Acc(%)	MAE	r
$\text{TFN}_{language}$	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

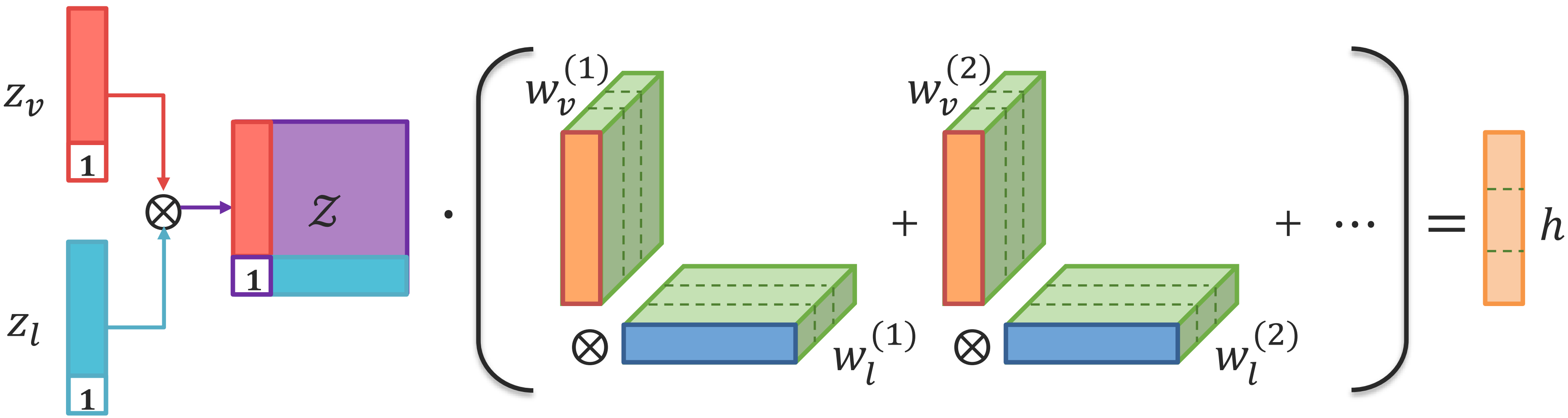
From Tensor Representation to Low-rank Fusion



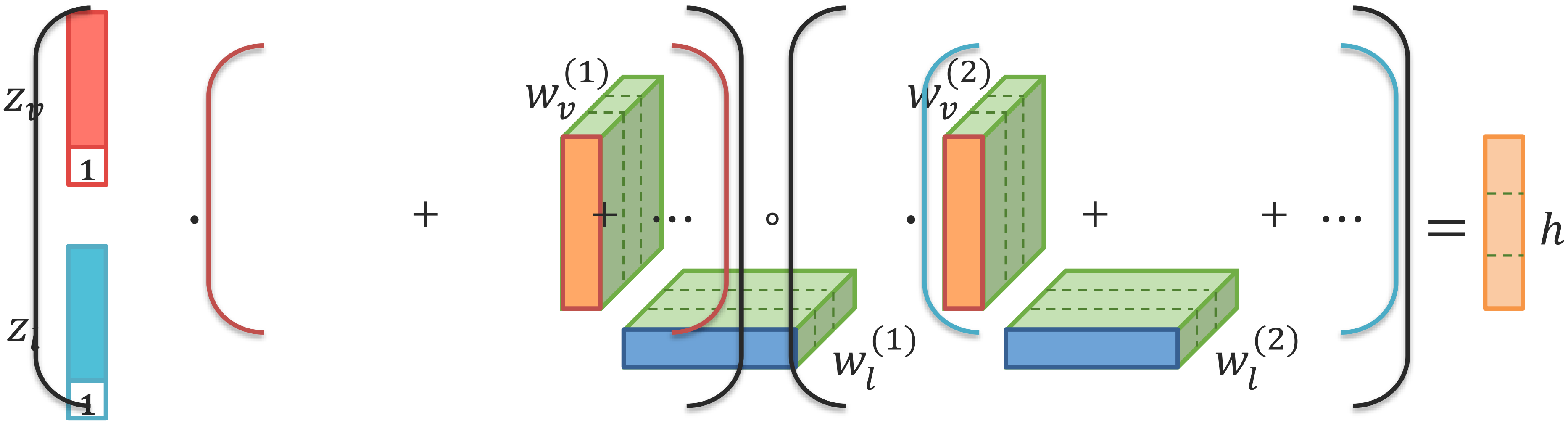
① Decomposition of weight tensor W



② Decomposition of Z



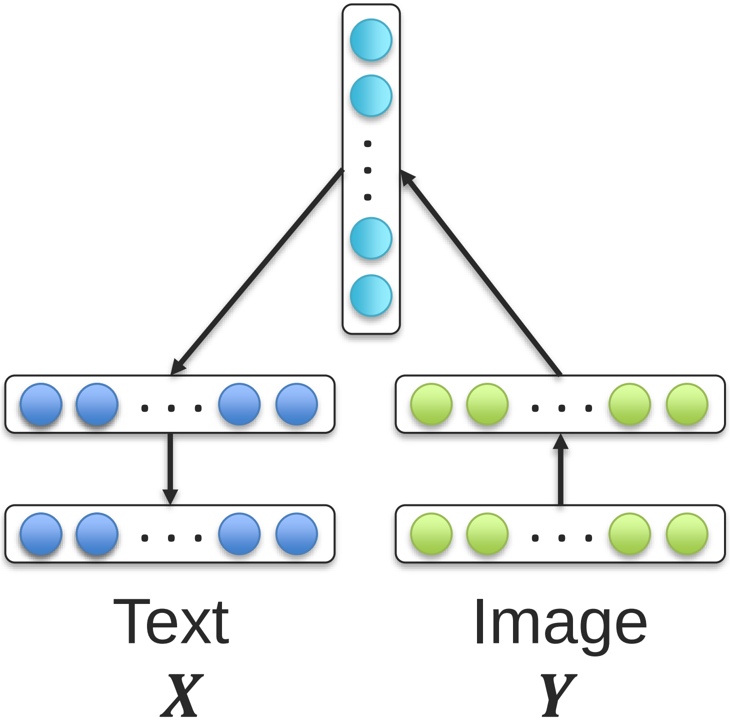
③ Rearranging computation



60

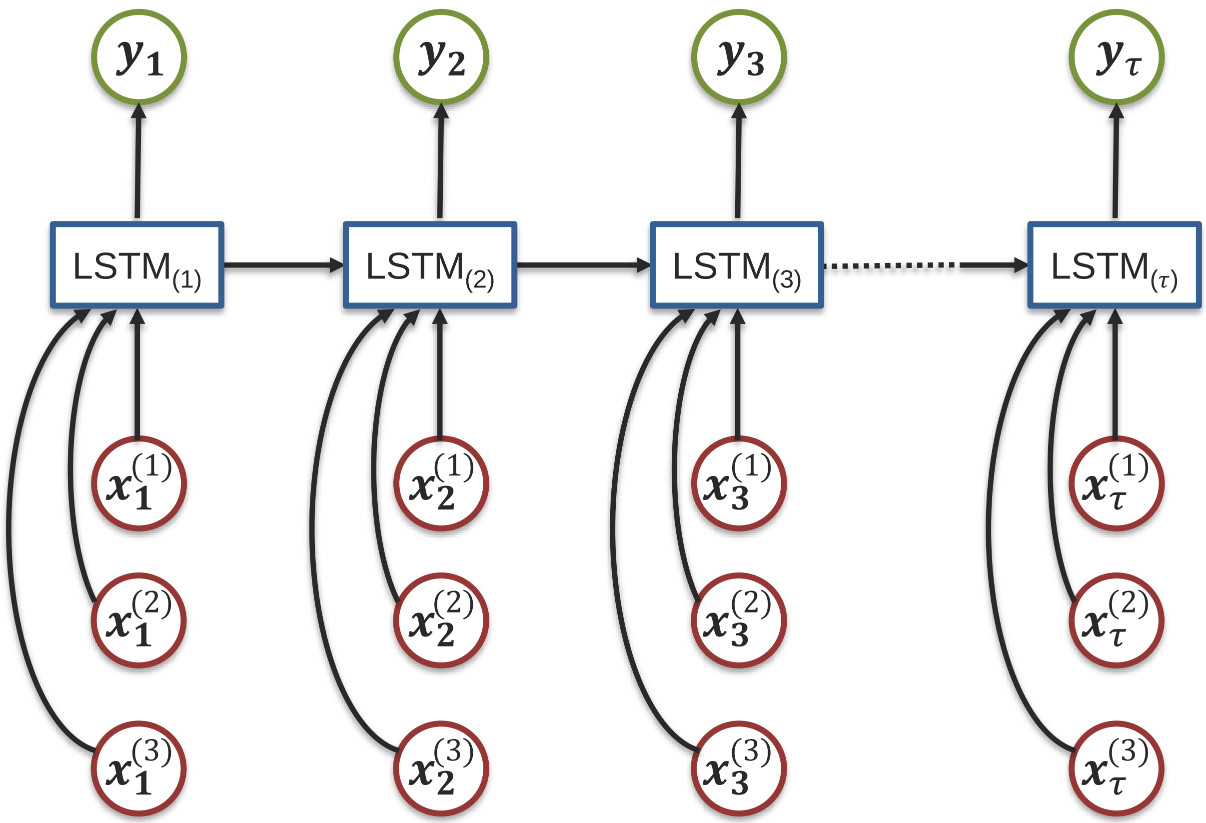
Multimodal Encoder-Decoder

- 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)

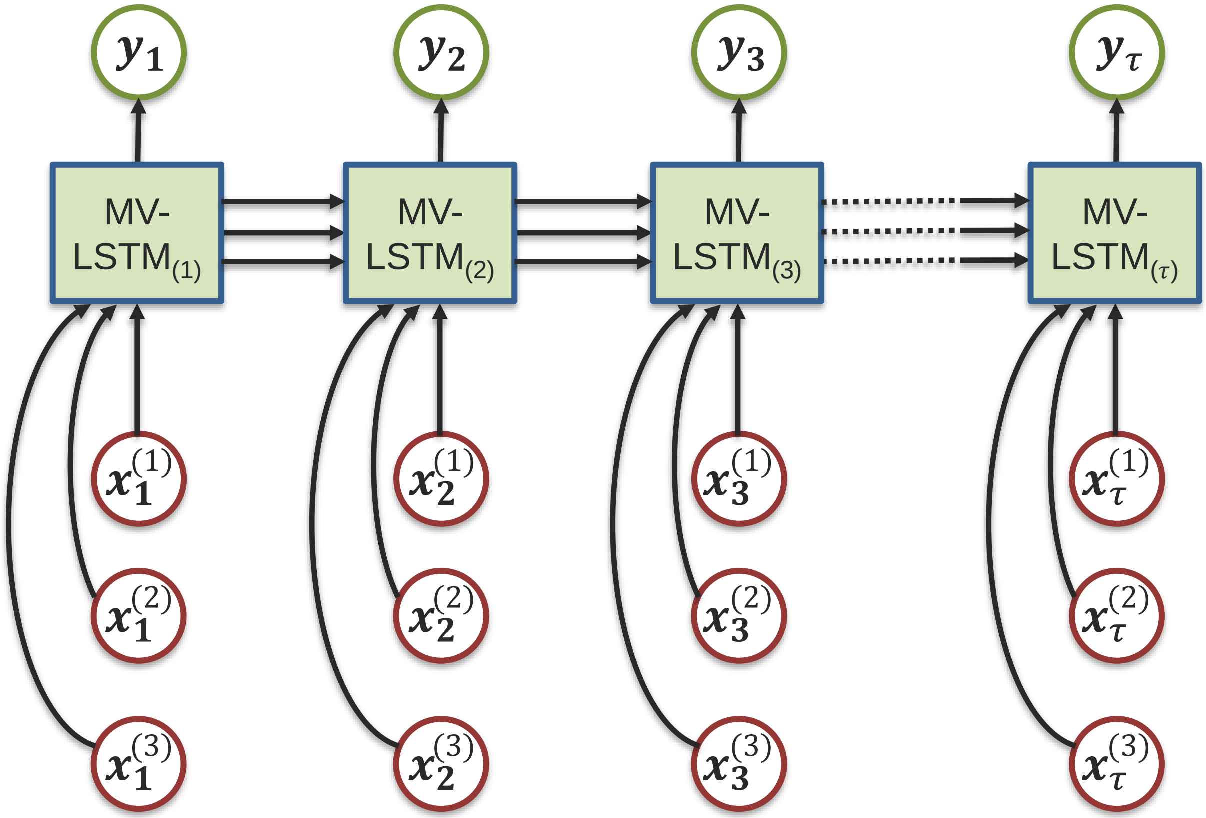


Multimodal LSTM

Multimodal Sequence Modeling – Early Fusion

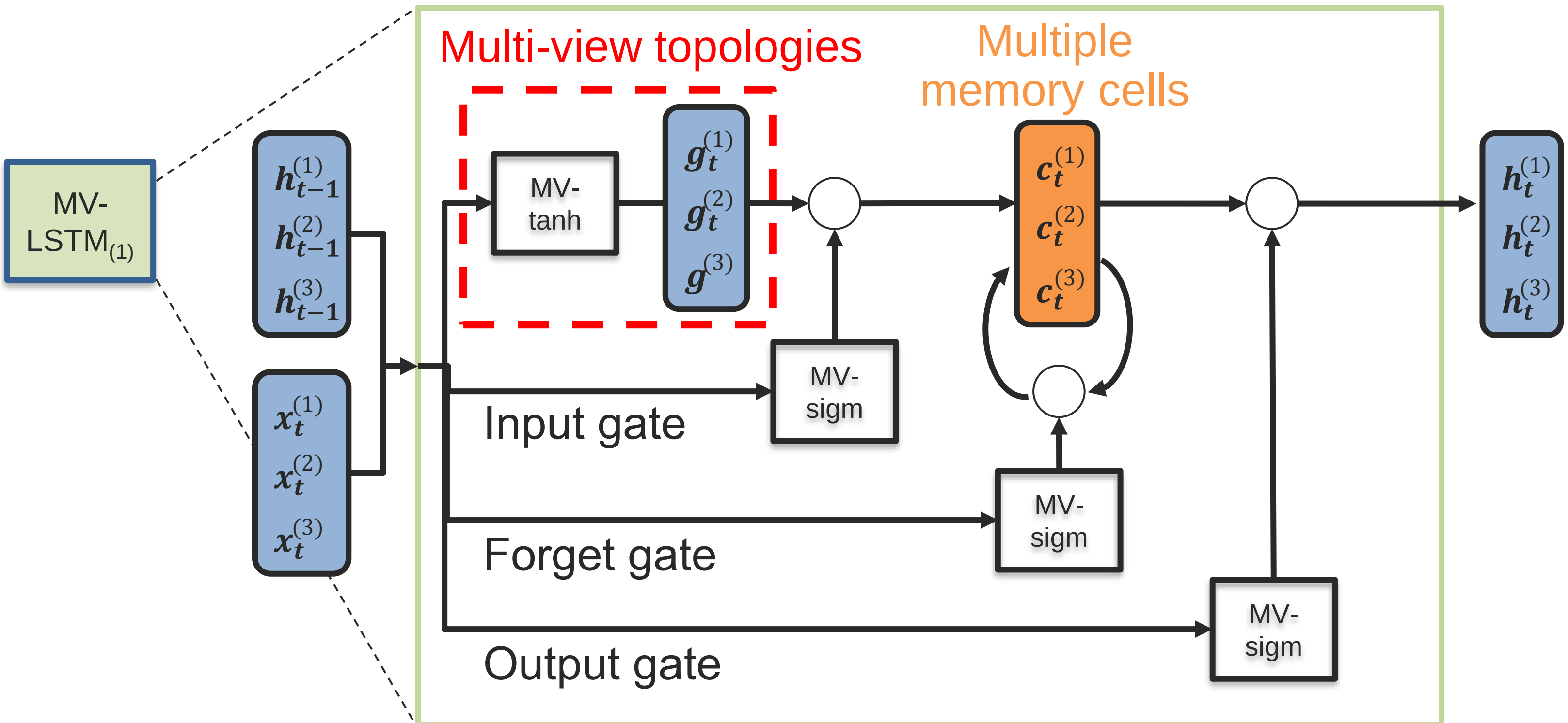


Multi-View Long Short-Term Memory (MV-LSTM)



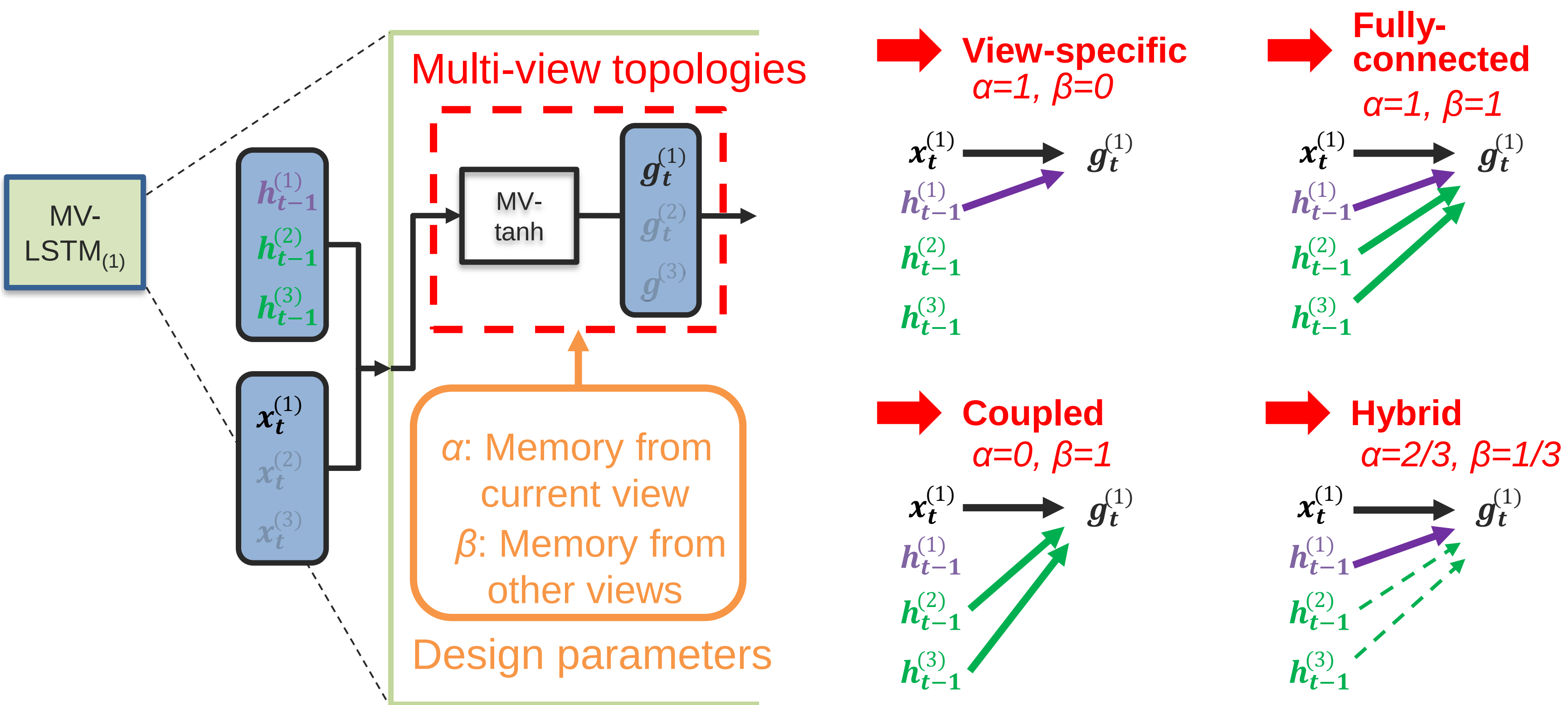
[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

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

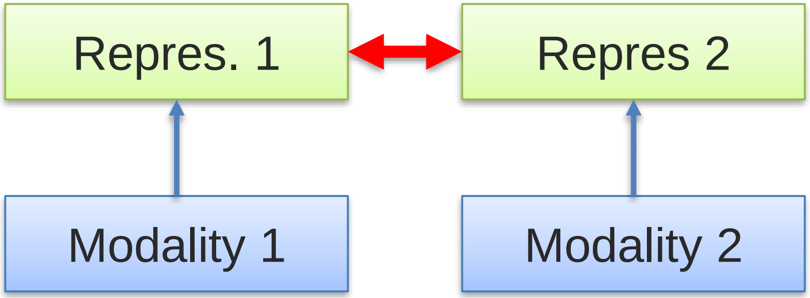
Class labels	Model	Precision	Recall	F1
Easy to engage	LSTM (Early fusion)	0.75	0.81	0.78
	MV-LSTM Full	0.81	0.81	0.81
	MV-LSTM Coupled	0.79	0.81	0.80
	MV-LSTM Hybrid	**0.80**	**0.86**	**0.83**
Difficult to engage	LSTM (Early fusion)	0.63	0.55	0.59
	MV-LSTM Full	0.68	0.68	0.68
	MV-LSTM Coupled	0.67	0.64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

Coordinated Multimodal Representations

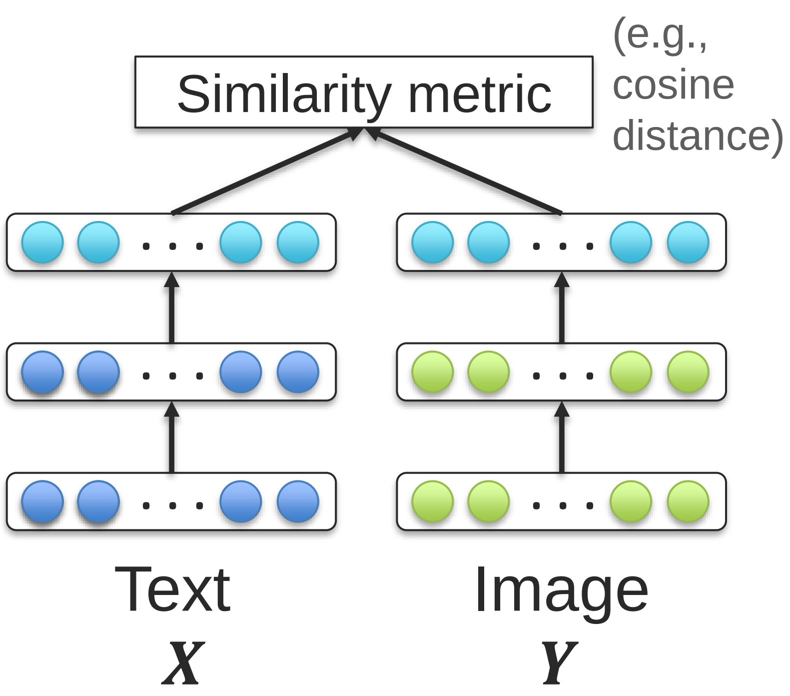
Coordinated multimodal embeddings

- Instead of projecting to a joint space enforce the similarity between unimodal embeddings



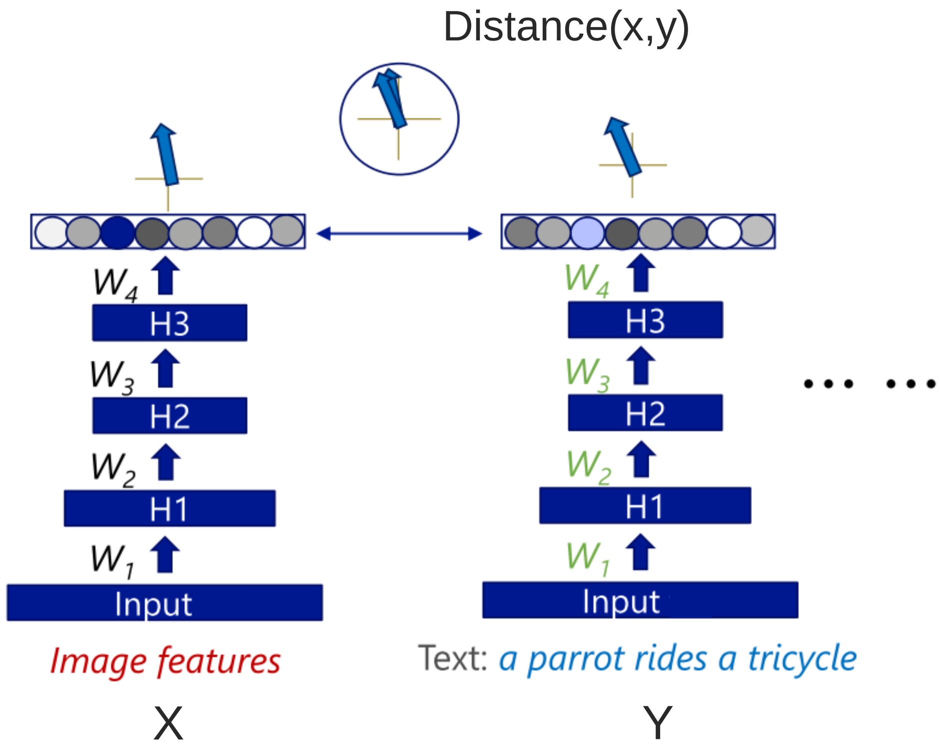
Coordinated Multimodal Representations

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



Coordinated Multimodal Embeddings

What should be the loss function?



[Frome et al., DeViSE: A Deep Visual-Semantic Embedding Model, NIPS 2013]

Max-Margin Loss – Multimodal Embeddings

What should be the loss function?

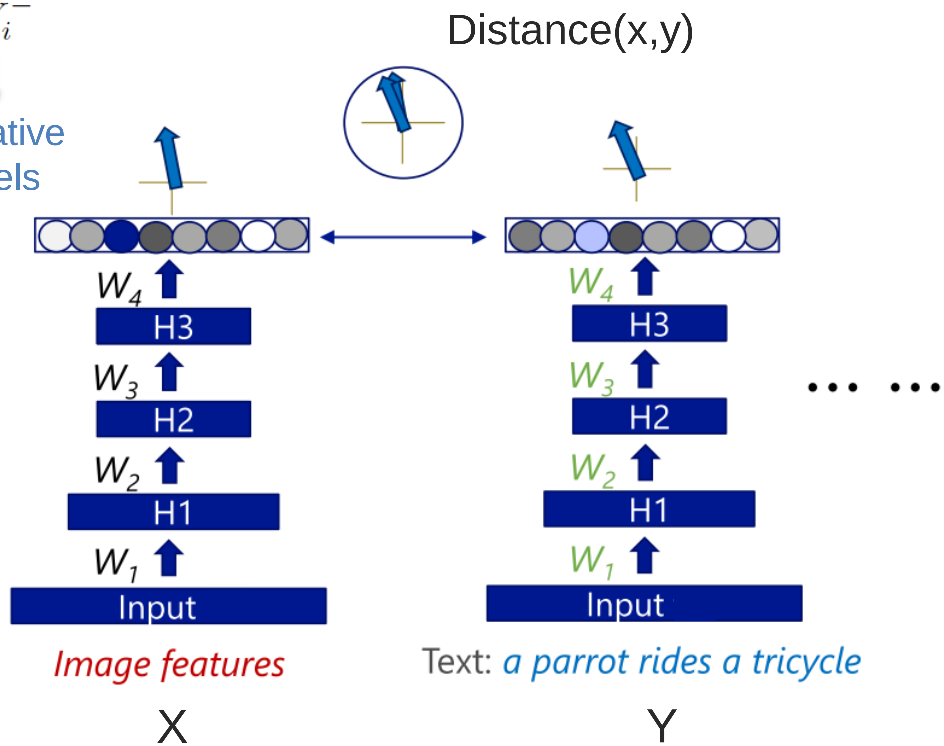
Max-margin:

$$d(x_i, y_j) + m < d(x_i, y_k) \quad \forall y_j \in Y_i^+, \forall y_k \in Y_i^-$$

Margin

Positive labels

Negative labels



[Frome et al., DeViSE: A Deep Visual-Semantic Embedding Model, NIPS 2013]

Structure-preserving Loss – Multimodal Embeddings

Symmetric max-margin:

$$d(x_i, y_j) + m < d(x_i, y_k) \quad \forall y_j \in Y_i^+, \forall y_k \in Y_i^-$$

$$d(x_{j'}, y_{i'}) + m < d(x_{k'}, y_{i'}) \quad \forall x_{j'} \in X_{i'}^+, \forall x_{k'} \in X_{i'}^-$$

+

Neighborhood of x_i:
images that share the
same meaning (text)

Structure-preserving constraints

$$d(x_i, x_j) + m < d(x_i, x_k) \quad \forall x_j \in N(x_i), \forall x_k \notin N(x_i)$$

$$d(y_{i'}, y_{j'}) + m < d(y_{i'}, y_{k'}) \quad \forall y_{j'} \in N(y_{i'}), \forall y_{k'} \notin N(y_{i'})$$

[Wang et al., Learning Deep Structure-Preserving Image-Text Embeddings, CVPR 2016]