Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 7.1: Multivariate Statistics

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Lecture Objectives

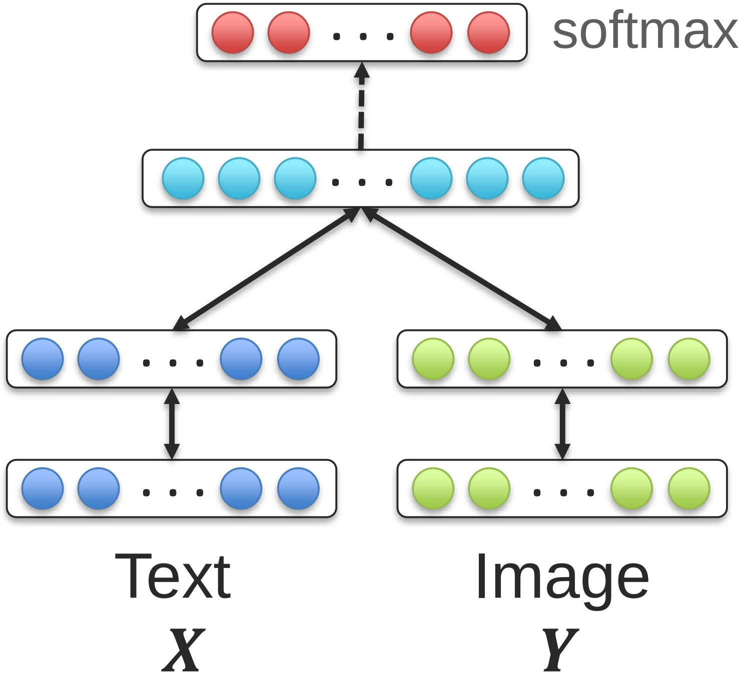
- Quick recap
  - Temporal joint representation
- Multivariate statistical analysis
  - Basic concepts (multivariate, covariance,…)
    - Principal component analysis (+SVD)
- Canonical Correlation Analysis
- Deep Correlation Networks
  - Deep CCA, DCCA-AutoEncoder
  - (Deep) Correlational neural networks
- Matrix Factorization
  - Nonnegative Matrix Factorization

Quick Recap

Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

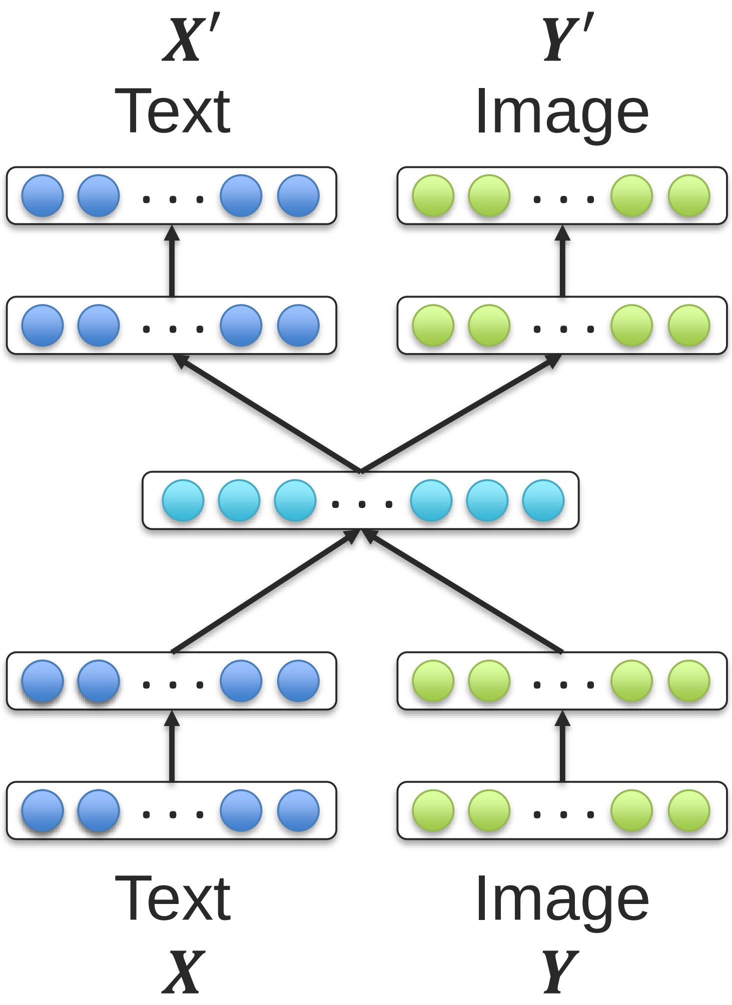
- ❑ Deep Multimodal Boltzmann machines



Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

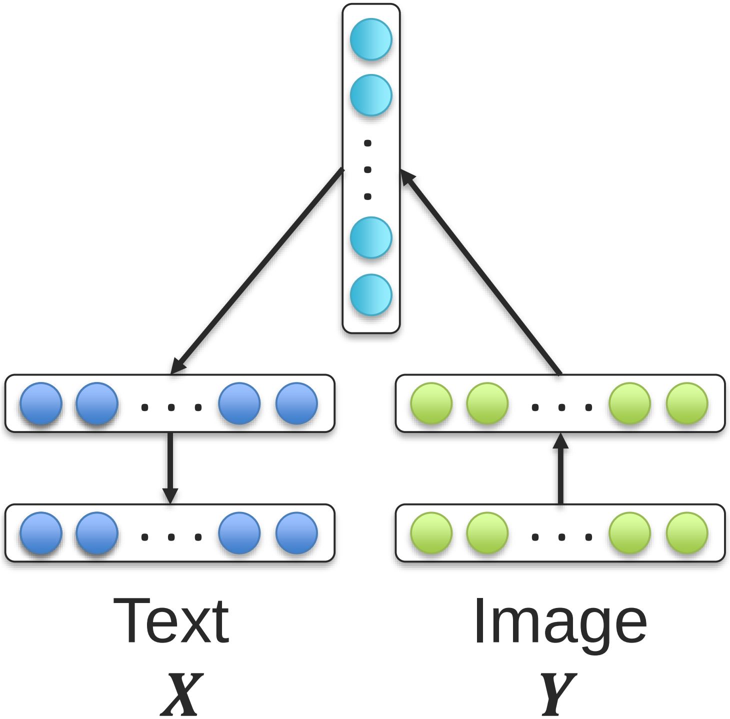
- ❑ Deep Multimodal Boltzmann machines
- ❑ Stacked Autoencoder



Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

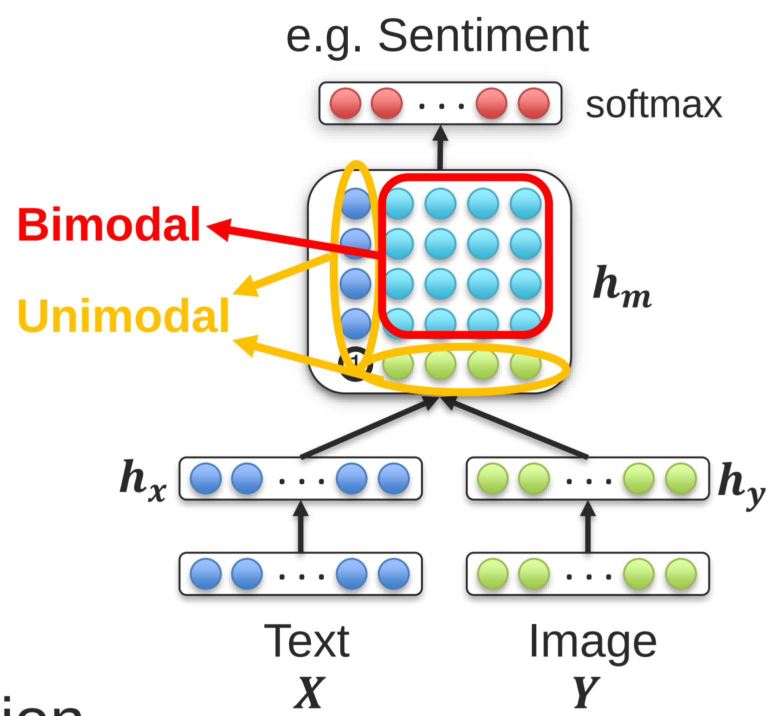
- ❑ Deep Multimodal Boltzmann machines
- ❑ Stacked Autoencoder
- ❑ Encoder-Decoder



Multimodal Representation Learning

Learn (unsupervised) a joint representation between multiple modalities where similar unimodal concepts are closely projected.

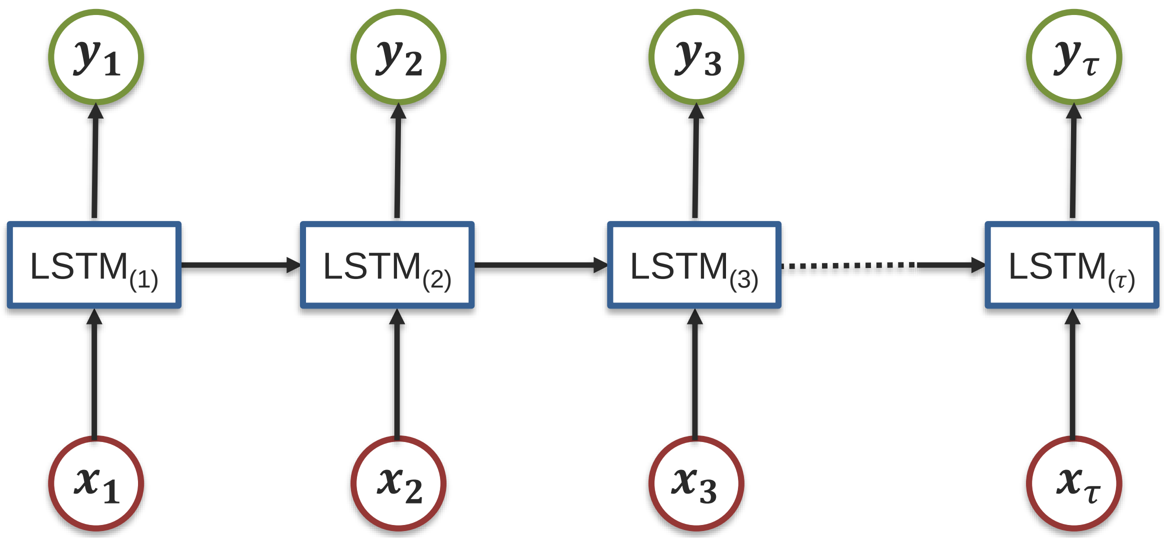
- ❑ Deep Multimodal Boltzmann machines
- ❑ Stacked Autoencoder
- ❑ Encoder-Decoder
- ❑ Tensor Fusion representation



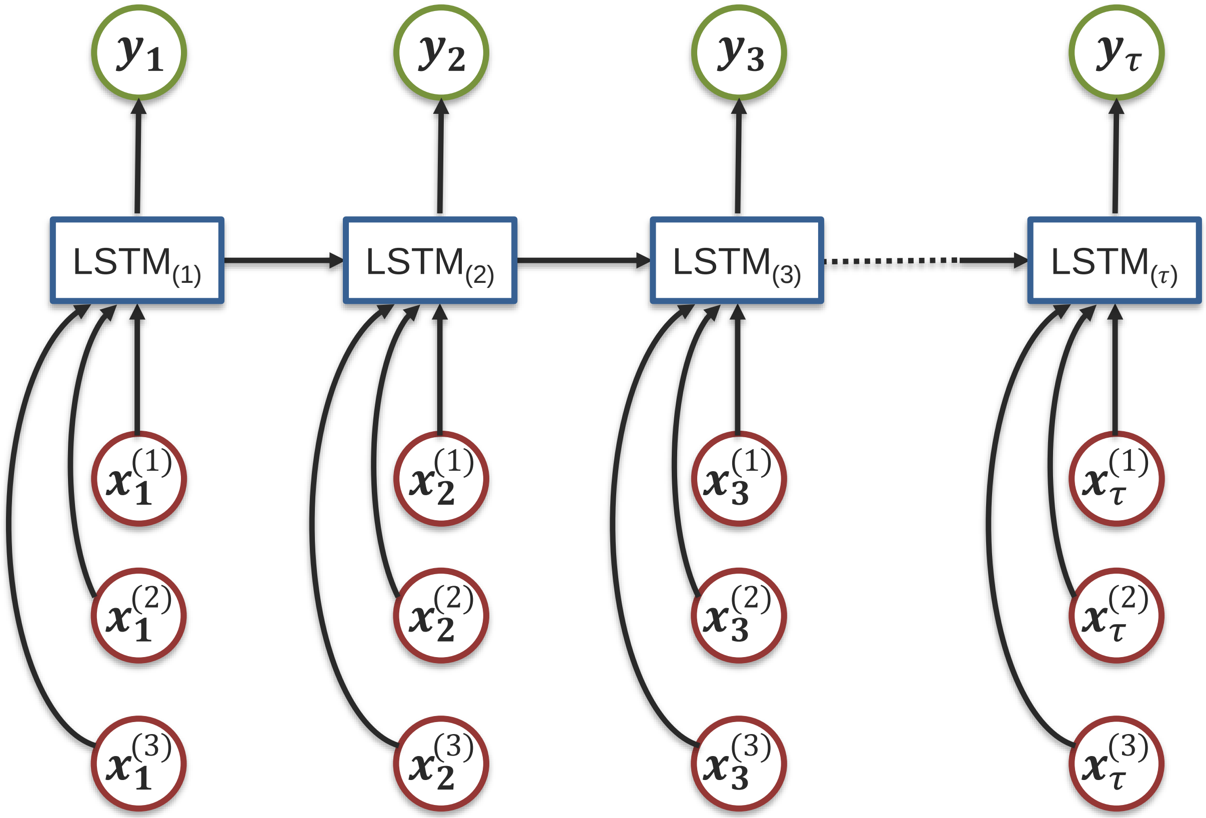
How Can We Learn Better Representations?

Temporal Joint Representation

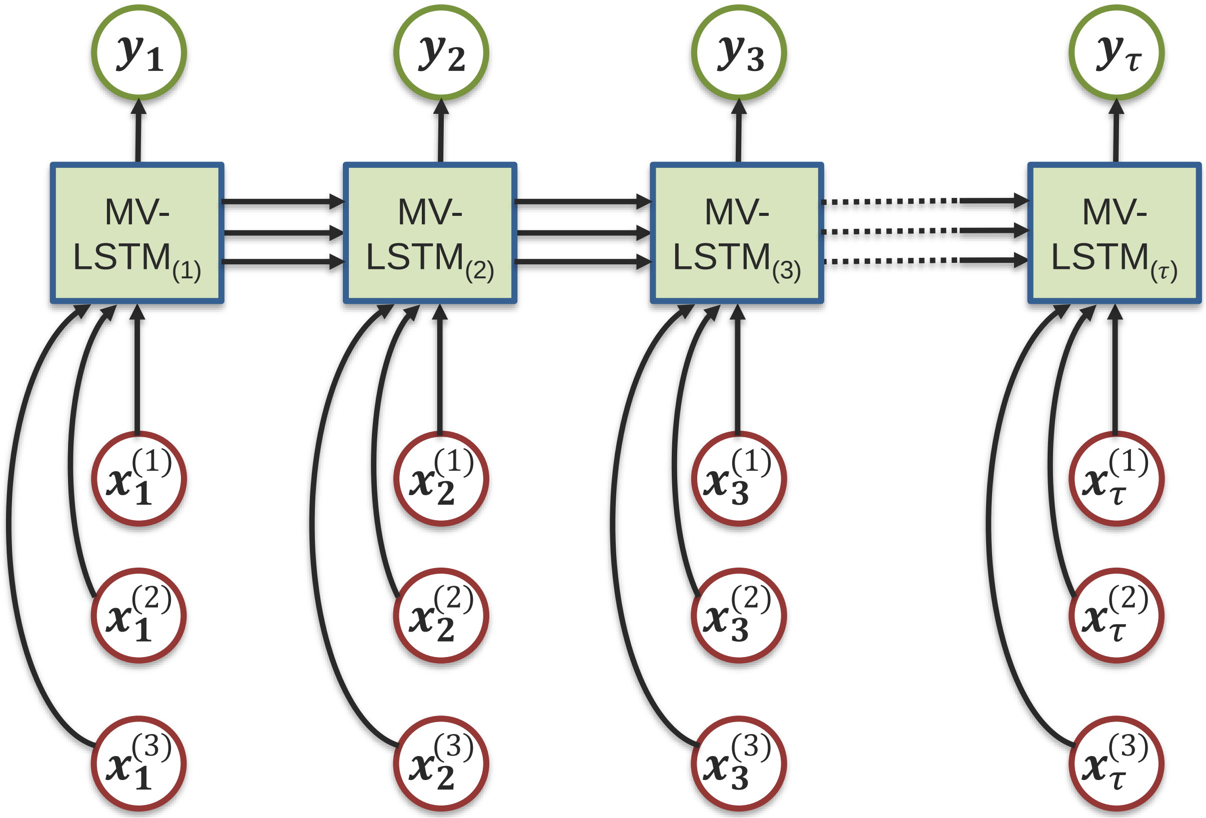
Sequence Representation with LSTM



Multimodal Sequence Representation – Early Fusion

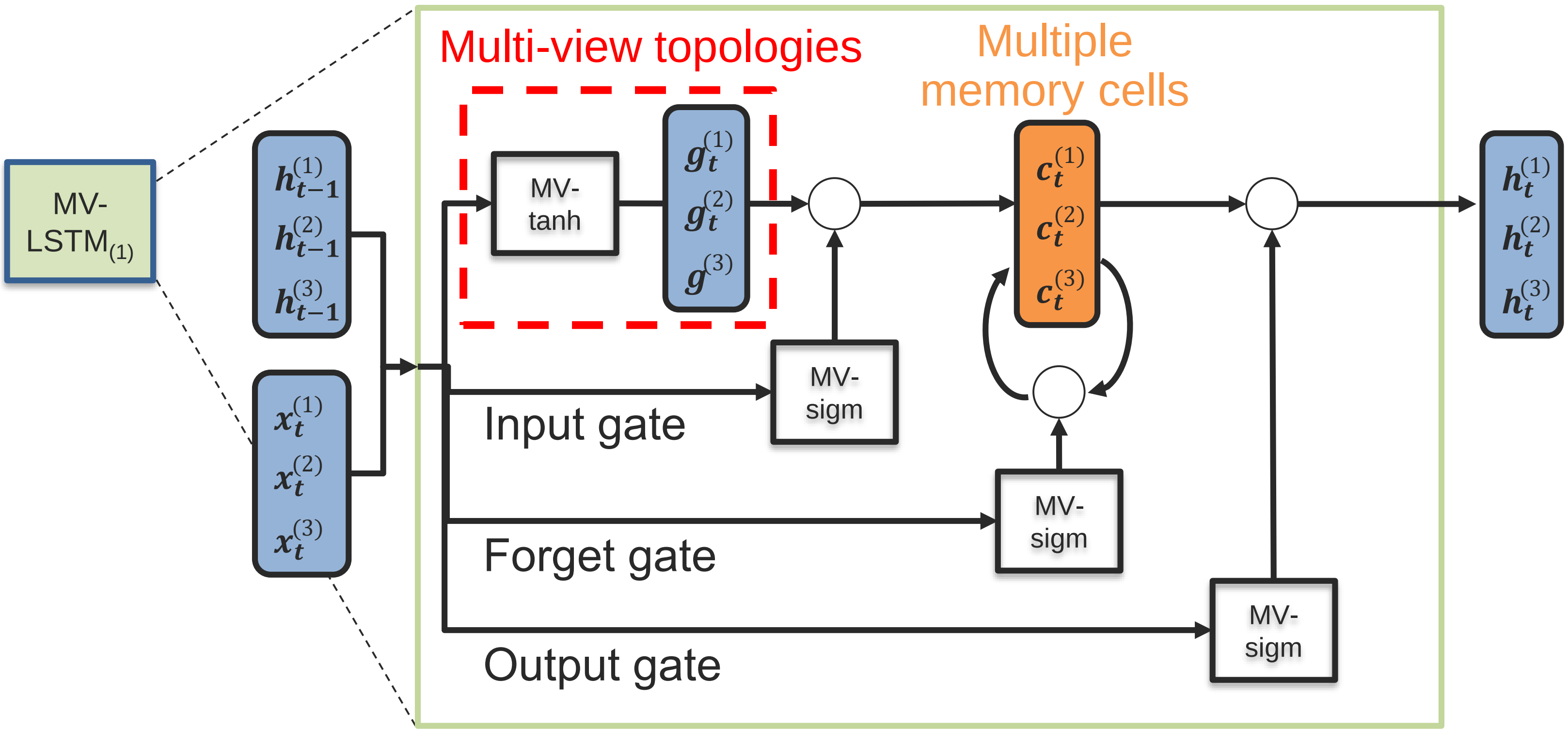


Multi-View Long Short-Term Memory (MV-LSTM)



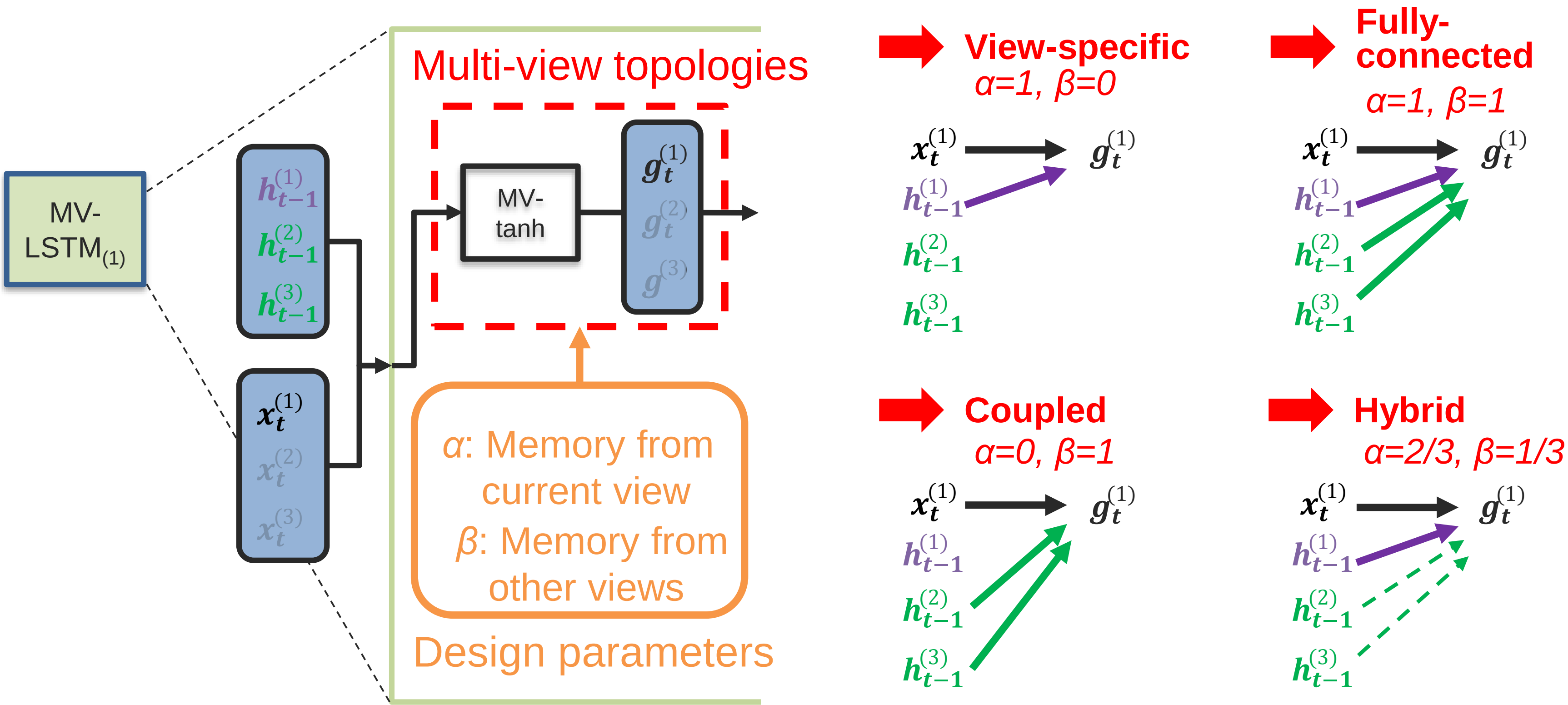
[Shyam, Morency, et al. Extending Long Short-Term Memory for Multi-View Structured Learning, **ECCV**, 2016]

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

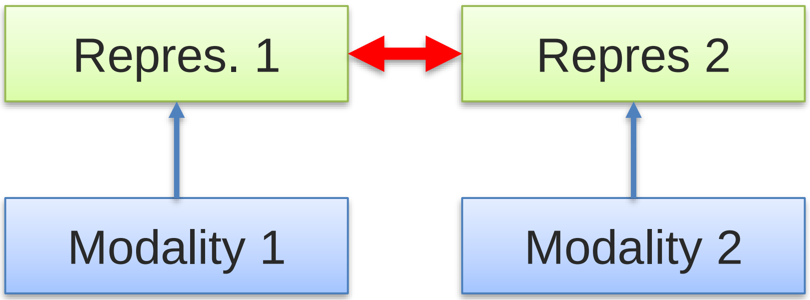
Class labels	Model	Precision	Recall	F1
Easy to engage	LSTM (Early fusion)	0.75	0.81	0.78
	MV-LSTM Full	0.81	0.81	0.81
	MV-LSTM Coupled	0.79	0.81	0.80
	MV-LSTM Hybrid	**0.80**	**0.86**	**0.83**
Difficult to engage	LSTM (Early fusion)	0.63	0.55	0.59
	MV-LSTM Full	0.68	0.68	0.68
	MV-LSTM Coupled	0.67	0.64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

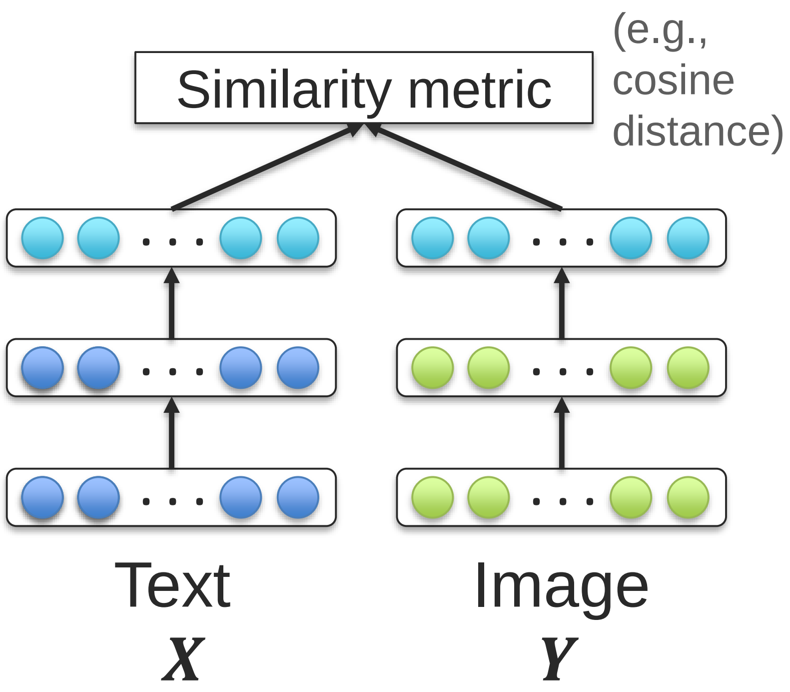
Coordinated multimodal embeddings

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

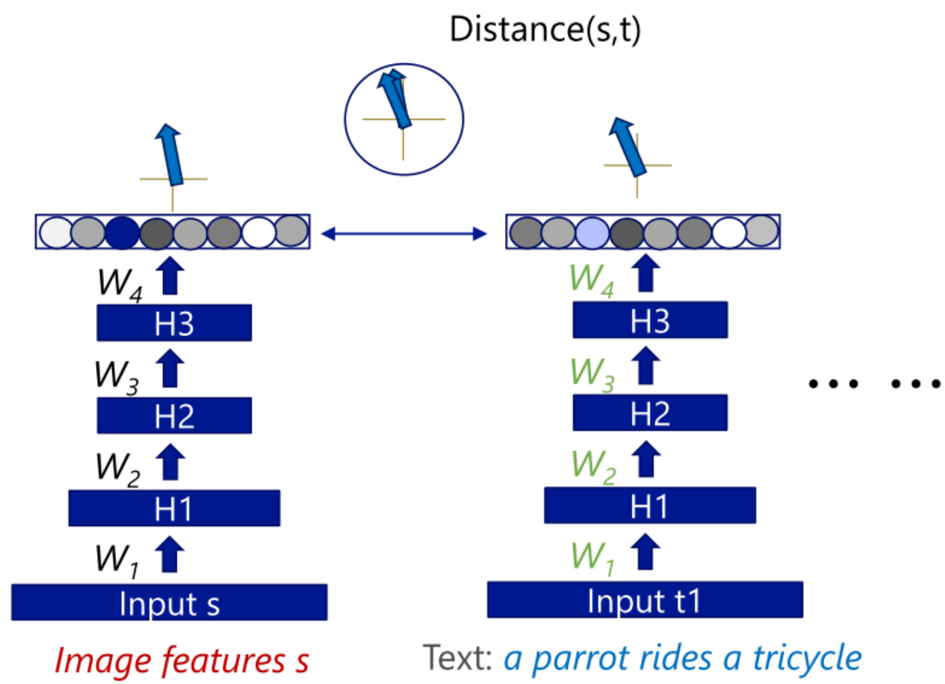


Coordinated Multimodal Representations

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

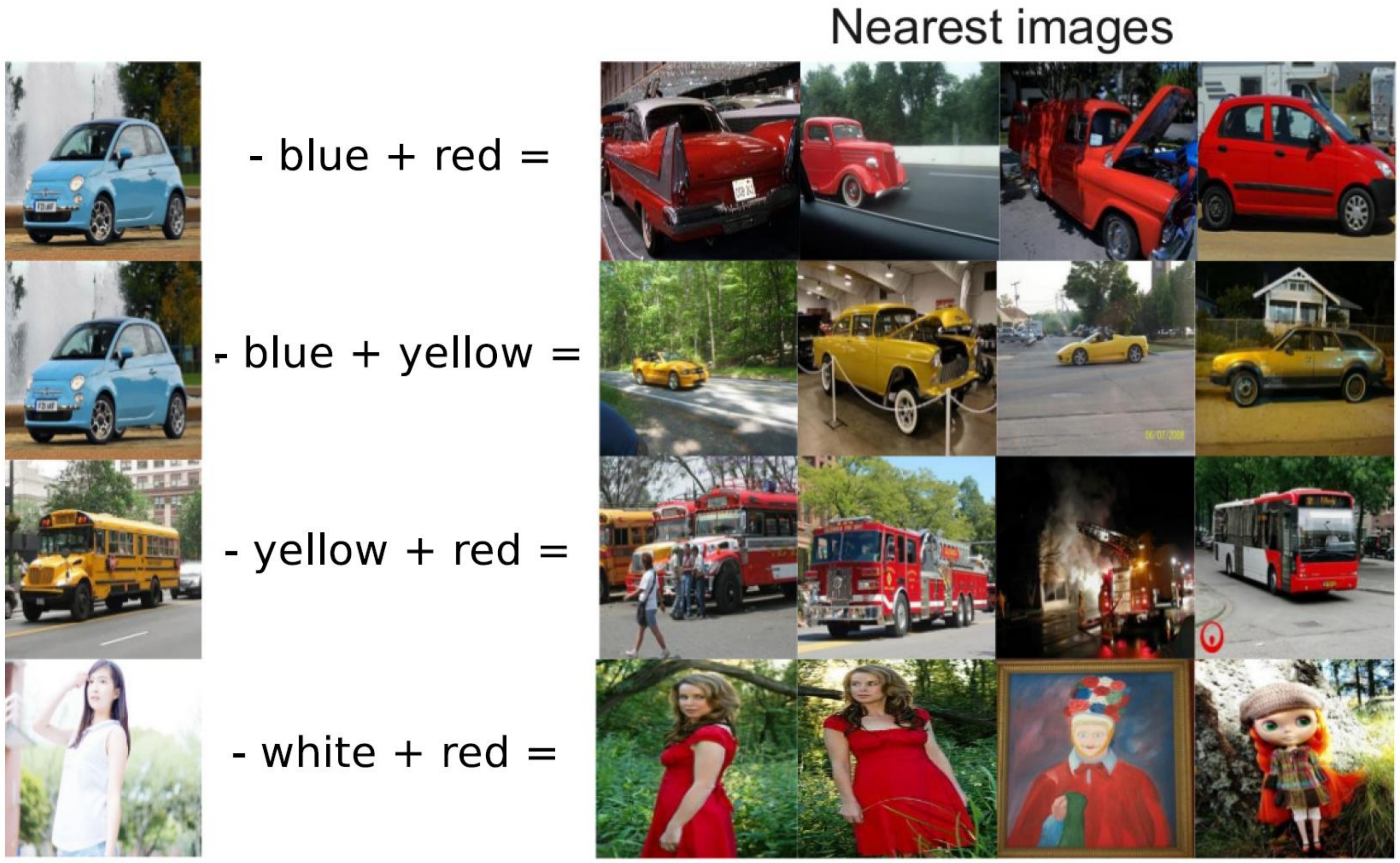


Coordinated Multimodal Embeddings



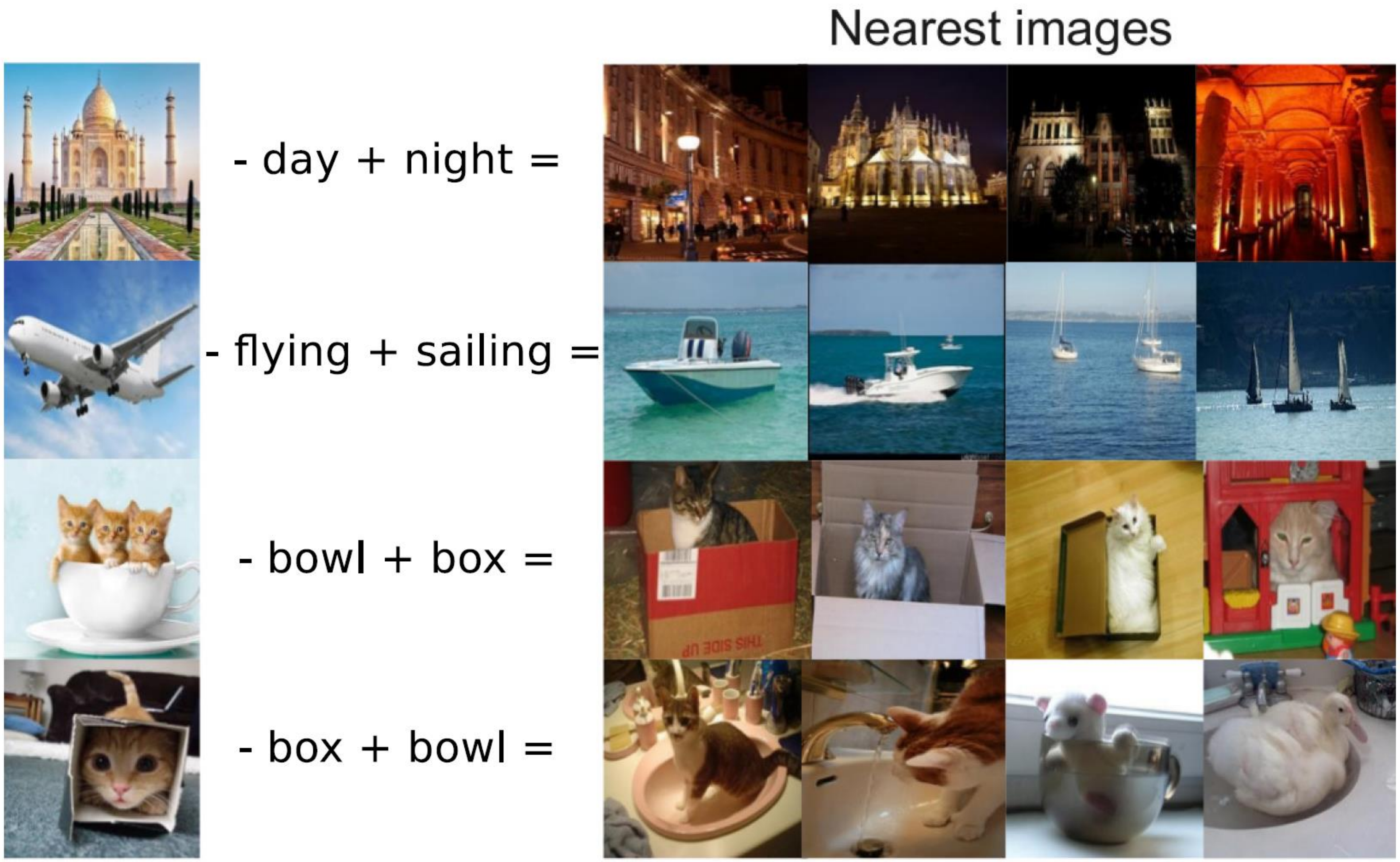
[Huang et al., Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, 2013]

Multimodal Vector Space Arithmetic



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

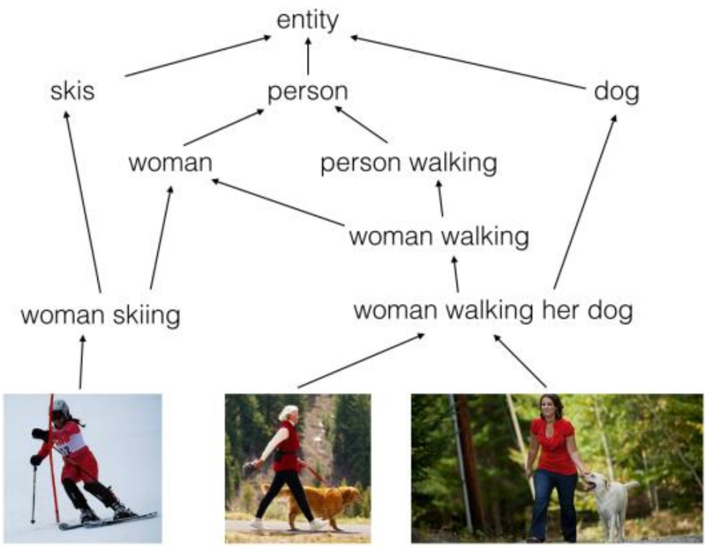
Multimodal Vector Space Arithmetic

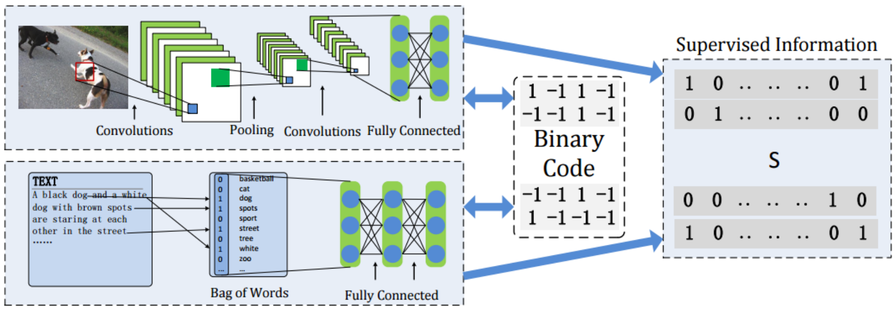


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

Structured coordinated embeddings

- Instead of or in addition to similarity add alternative structure





[Vendrov et al., Order-Embeddings of Images and Language, 2016]

[Jiang and Li, Deep Cross-Modal Hashing]

Multivariate Statistical Analysis

Multivariate Statistical Analysis

"Statistical approaches to understand the relationships in high dimensional data"

- Example of multivariate analysis approaches:
  - Multivariate analysis of variance (MANOVA)
  - Principal components analysis (PCA)
  - Factor analysis
  - Linear discriminant analysis (LDA)
  - Canonical correlation analysis (CCA)

Random Variables

Definition: A variable whose possible values are numerical outcomes of a random phenomenon.

- ❑ **Discrete** random variable is one which may take on only a countable number of distinct values such as 0,1,2,3,4,…
- ❑ **Continuous** random variable is one which takes an infinite number of possible values.

Examples of random variables:

- Someone’s age
- Someone’s height
- Someone’s weight

Discrete or continuous?

Correlated?

Definitions

Given two random variables X and Y:

Expected value probability-weighted average of all possible values

$$\mu = E[X] = \sum_i x_i P(x_i)$$

- If same probability for all observations x_i, then same as arithmetic mean

Variance measures the spread of the observations

$$\sigma^2 = Var(X) = E[(X-\mu)(X-\mu)] = E[\bar{X}\bar{X}]$$

If data is centered

- Variance is equal to the square of the standard deviation σ

Covariance measures how much two random variables change together

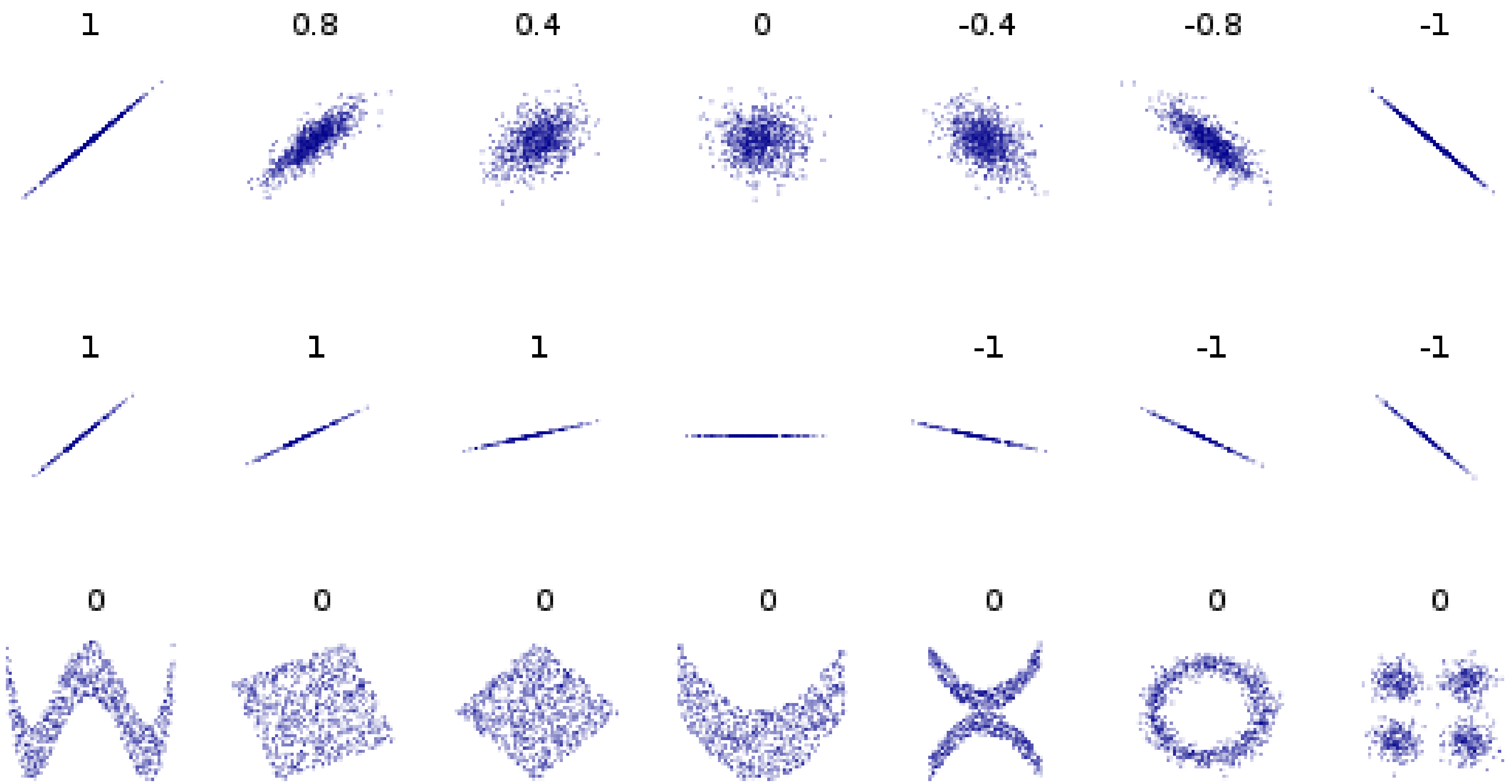
$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_y)] = E[\bar{X}\bar{Y}]$$

Definitions

Pearson Correlation measures the extent to which two variables have a linear relationship with each other

$$\rho_{X,Y} = corr(X, Y) = \frac{cov(X, Y)}{var(X)var(Y)}$$

Pearson Correlation Examples



Definitions

Multivariate (multidimensional) random variables

(aka random vector)

$$\boldsymbol{X} = [X^1, X^2, X^3, \dots, X^M]$$

$$\boldsymbol{Y} = [Y^1, Y^2, Y^3, \dots, Y^N]$$

Covariance matrix generalizes the notion of variance

$$\Sigma_{\boldsymbol{X}} = \Sigma_{\boldsymbol{X},\boldsymbol{X}} = var(\boldsymbol{X}) = E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T] = E[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T]$$

Cross-covariance matrix generalizes the notion of covariance

$$\Sigma_{\boldsymbol{X},\boldsymbol{Y}} = cov(\boldsymbol{X}, \boldsymbol{Y}) = E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{Y} - E[\boldsymbol{Y}])^T] = E[\overline{\boldsymbol{X}}\overline{\boldsymbol{Y}}^T]$$

Definitions

Multivariate (multidimensional) random variables

(aka random vector)

$$\boldsymbol{X} = [X^1, X^2, X^3, \dots, X^M]$$

$$\boldsymbol{Y} = [Y^1, Y^2, Y^3, \dots, Y^N]$$

Covariance matrix generalizes the notion of variance

$$\Sigma_{\boldsymbol{X}} = \Sigma_{\boldsymbol{X},\boldsymbol{X}} = var(\boldsymbol{X}) = E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])^T] = E[\overline{\boldsymbol{X}}\overline{\boldsymbol{X}}^T]$$

Cross-covariance matrix generalizes the notion of covariance

$$\Sigma_{\boldsymbol{X},\boldsymbol{Y}} = cov(\boldsymbol{X}, \boldsymbol{Y}) = \begin{bmatrix} cov(X_1, Y_1) & cov(X_2, Y_1) & \cdots & cov(X_M, Y_1) \\ cov(X_1, Y_2) & cov(X_2, Y_2) & \cdots & cov(X_M, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ cov(X_1, Y_N) & cov(X_2, Y_N) & \dots & cov(X_M, Y_N) \end{bmatrix}$$

Definitions – Matrix Operations

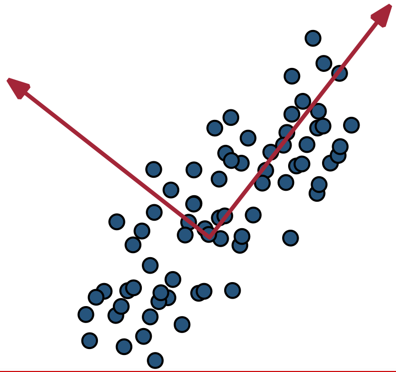
Trace is defined as the sum of the elements on the main diagonal of any matrix $\boldsymbol{X}$

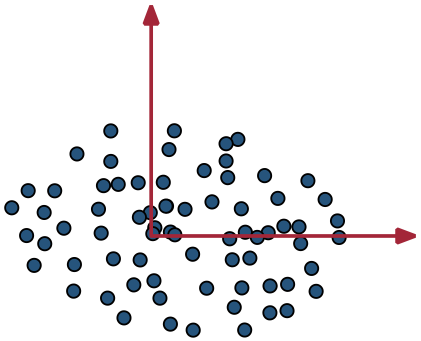
$$tr(\boldsymbol{X}) = \sum_{i=1}^{n} x_{ii}$$

Principal component analysis

PCA converts a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called *principal components*

- Eigenvectors are orthogonal towards each other and have length one
- The first couple of eigenvectors explain the most of the variance observed in the data
- Low eigenvalues indicate little loss of information if omitted





Eigenvalues and Eigenvectors

Eigenvalue decomposition

If A is an $n \times n$ matrix, do there exist nonzero vectors $\mathbf{x}$ in R^n such that $A\mathbf{x}$ is a scalar multiple of $\mathbf{x}$?

- (The term eigenvalue is from the German word *Eigenwert*, meaning "proper value")

Eigenvalue equation:

$$A\mathbf{x} = \lambda \mathbf{x}$$

Eigenvector

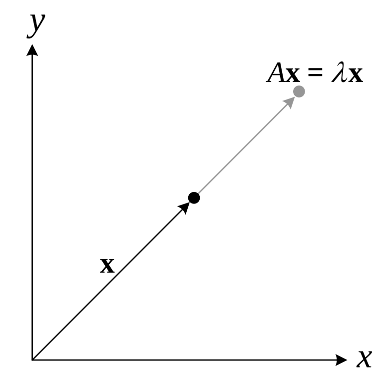
Eigenvalue

A: an $n \times n$ matrix

λ: a scalar (could be **zero**)

$\mathbf{x}$: a **nonzero** vector in R^n

Geometric Interpretation



Singular Value Decomposition (SVD)

- SVD expresses any matrix $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

- The columns of $\mathbf{U}$ are eigenvectors of $\mathbf{A}\mathbf{A}^T$, and the columns of $\mathbf{V}$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$.

$$\mathbf{A}\mathbf{A}^T\mathbf{u}_i = s_i^2\mathbf{u}_i$$
$$\mathbf{A}^T\mathbf{A}\mathbf{v}_i = s_i^2\mathbf{v}_i$$

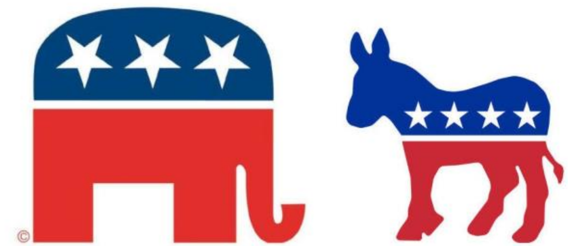
Canonical Correlation Analysis

Multi-view Learning

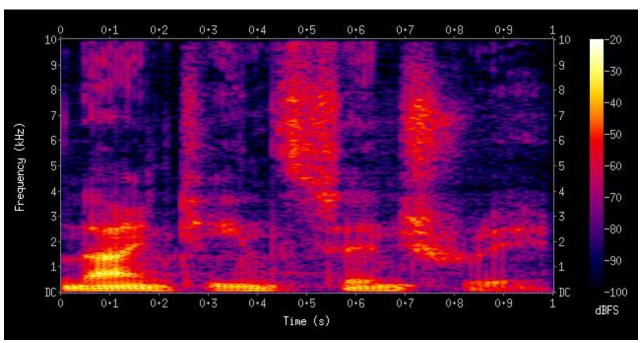
X

demographic properties

Y



responses to survey



audio features at time i

video features at time i

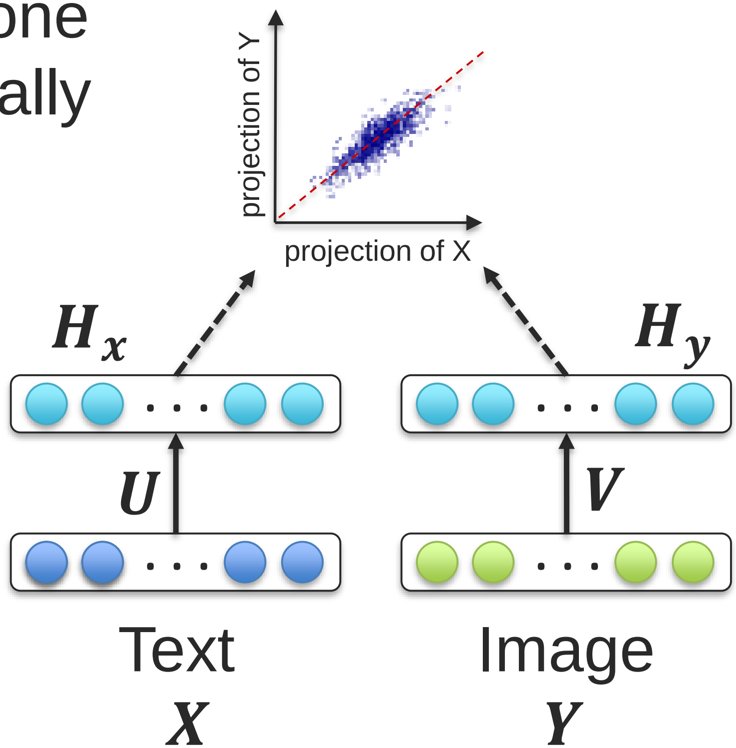
Canonical Correlation Analysis

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

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

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

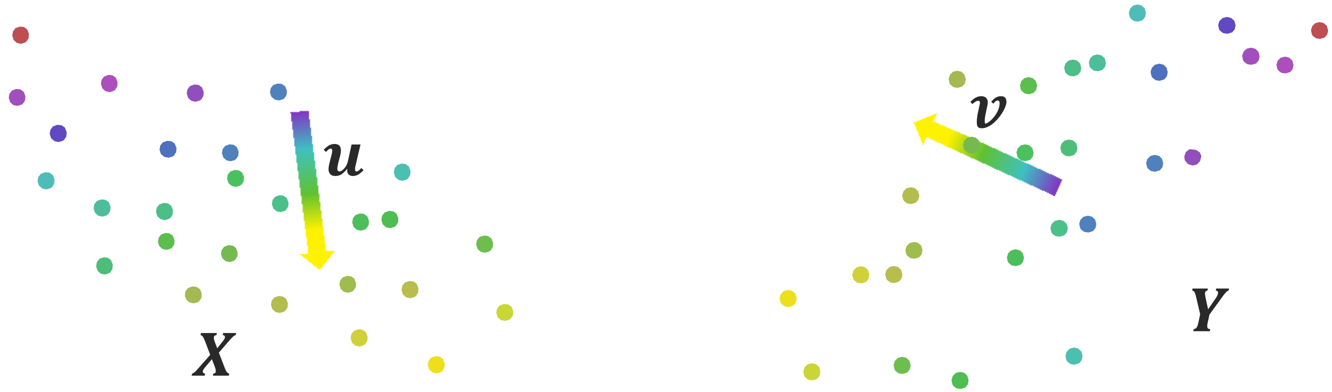
$$= \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr(\boldsymbol{u}^T\boldsymbol{X}, \boldsymbol{v}^T\boldsymbol{Y})$$



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

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

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

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

$$= \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{Y}^T\boldsymbol{v}}{\sqrt{\boldsymbol{u}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{u}}\sqrt{\boldsymbol{v}^T\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{v}}}$$

$$= \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}} \frac{\boldsymbol{u}^T\boldsymbol{\Sigma}_{\boldsymbol{XY}}\boldsymbol{v}}{\sqrt{\boldsymbol{u}^T\boldsymbol{\Sigma}_{\boldsymbol{XX}}\boldsymbol{u}}\sqrt{\boldsymbol{v}^T\boldsymbol{\Sigma}_{\boldsymbol{YY}}\boldsymbol{v}}}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{XY}} = cov(\boldsymbol{X}, \boldsymbol{Y}) = \boldsymbol{X}\boldsymbol{Y}^T$$

if both $\boldsymbol{X}, \boldsymbol{Y}$ have 0 mean

$$\boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{0} \qquad \boldsymbol{\mu}_{\boldsymbol{Y}} = \boldsymbol{0}$$

Canonical Correlation Analysis

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

$$\left(\boldsymbol{u}_{(i)}^{*}, \boldsymbol{v}_{(i)}^{*}\right) = \underset{\boldsymbol{u}_{(i)}, \boldsymbol{v}_{(i)}}{\operatorname{argmax}} \frac{\boldsymbol{u}_{(i)}^{T} \boldsymbol{\Sigma}_{\boldsymbol{XY}} \boldsymbol{v}_{(i)}}{\sqrt{\boldsymbol{u}_{(i)}^{T} \boldsymbol{\Sigma}_{\boldsymbol{XX}} \boldsymbol{u}_{(i)}} \sqrt{\boldsymbol{v}_{(i)}^{T} \boldsymbol{\Sigma}_{\boldsymbol{YY}} \boldsymbol{v}_{(i)}}}$$

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

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

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

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

Canonical Correlation Analysis

$$(\boldsymbol{U}^*, \boldsymbol{V}^*) = \underset{\boldsymbol{U},\boldsymbol{V}}{\operatorname{argmax}} \frac{tr(\boldsymbol{U}^T \boldsymbol{\Sigma}_{\boldsymbol{XY}} \boldsymbol{V})}{\sqrt{\boldsymbol{U}^T \boldsymbol{\Sigma}_{\boldsymbol{XX}} \boldsymbol{U}} \sqrt{\boldsymbol{V}^T \boldsymbol{\Sigma}_{\boldsymbol{YY}} \boldsymbol{V}}}$$

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

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

Canonical Correlation Analysis:

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

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

Canonical Correlation Analysis

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

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

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

Canonical Correlation Analysis

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

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

How to solve it? ➢ Lagrange Multipliers!

Lagrange function

$$\boldsymbol{L} = tr(\boldsymbol{U}^T \boldsymbol{\Sigma}_{XY} \boldsymbol{V}) + \alpha\left(\boldsymbol{U}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{U} - \boldsymbol{I}\right) + \beta(\boldsymbol{V}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{V} - \boldsymbol{I})$$

➢ And then find stationary points of L: $\frac{\partial L}{\partial \boldsymbol{U}} = 0 \quad \frac{\partial L}{\partial \boldsymbol{V}} = 0$

$$\boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{XY}^{T} \boldsymbol{U} = \boldsymbol{\lambda} \boldsymbol{U}$$

$$\boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{XY}^{T} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{V} = \boldsymbol{\lambda} \boldsymbol{V} \quad \text{where } \lambda = 4\alpha\beta$$

Canonical Correlation Analysis

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

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

$$\boldsymbol{T} \triangleq \boldsymbol{\Sigma}_{XX}^{-1/2} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-1/2}$$

$$(\boldsymbol{U}^*, \boldsymbol{V}^*) = (\boldsymbol{\Sigma}_{XX}^{-1/2} \boldsymbol{U}_{SVD}, \boldsymbol{\Sigma}_{YY}^{-1/2} \boldsymbol{V}_{SVD})$$

- ➢ Can solve these eigenvalue equations with Singular Value Decomposition (SVD)

Eigenvalues

Eigenvectors

Eigenvalue equations

$$\begin{cases} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{XY}^{T} \boldsymbol{U} = \boldsymbol{\lambda} \boldsymbol{U} \\ \boldsymbol{\Sigma}_{YY}^{-1} \boldsymbol{\Sigma}_{XY}^{T} \boldsymbol{\Sigma}_{XX}^{-1} \boldsymbol{\Sigma}_{XY} \boldsymbol{V} = \boldsymbol{\lambda} \boldsymbol{V} \end{cases}$$

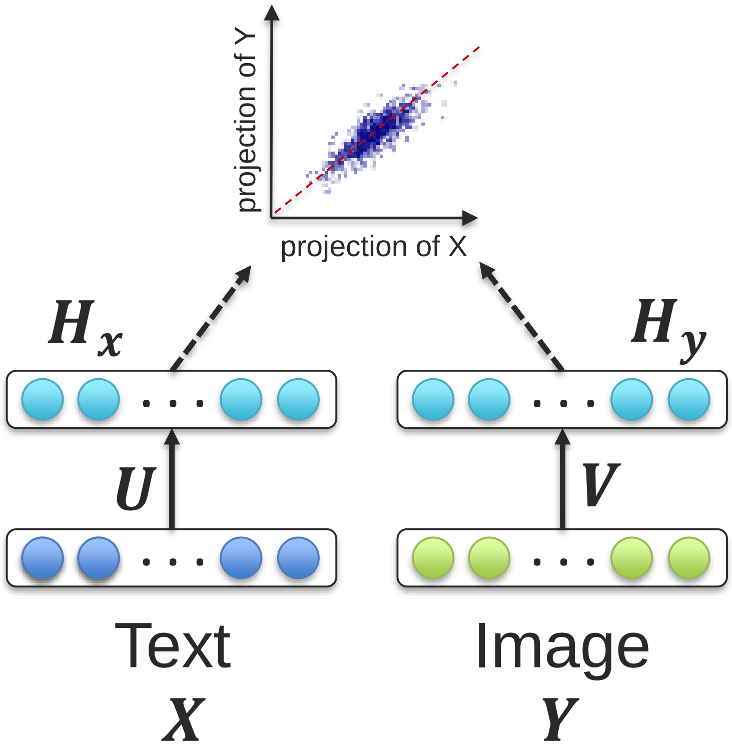
where $\lambda = 4\alpha\beta$

Canonical Correlation Analysis

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

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

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



Exploring Deep Correlation Networks

Deep Canonical Correlation Analysis

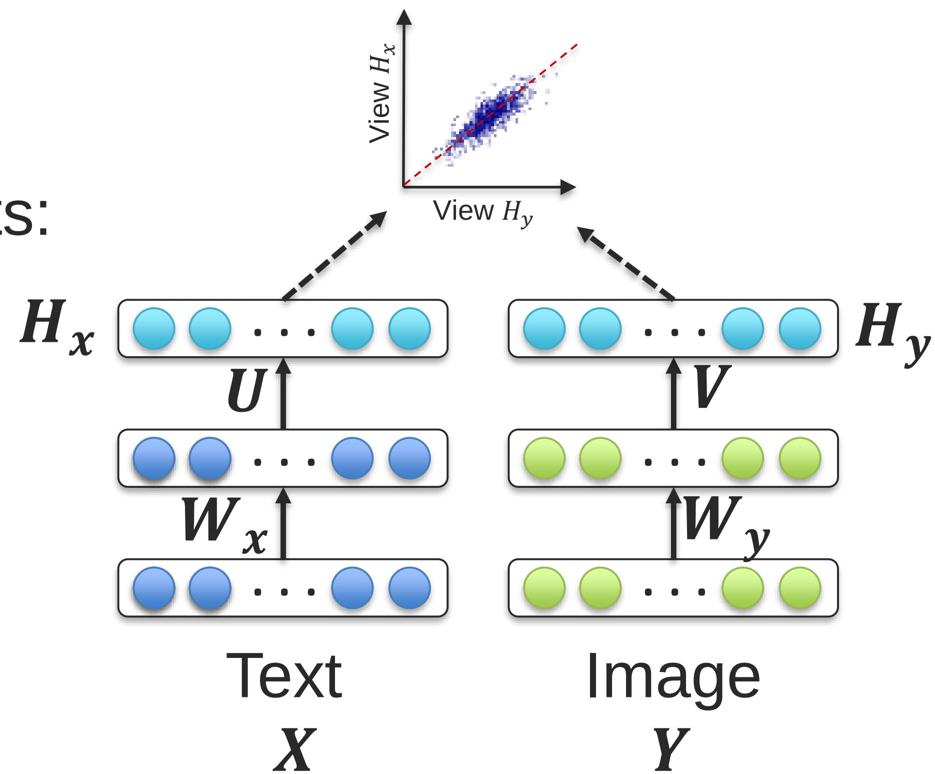
Same objective function as CCA:

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

And need to compute gradients:

$$\frac{\partial corr(\boldsymbol{H_x}, \boldsymbol{H_y})}{\partial U}$$

$$\frac{\partial corr(\boldsymbol{H_x}, \boldsymbol{H_y})}{\partial V}$$

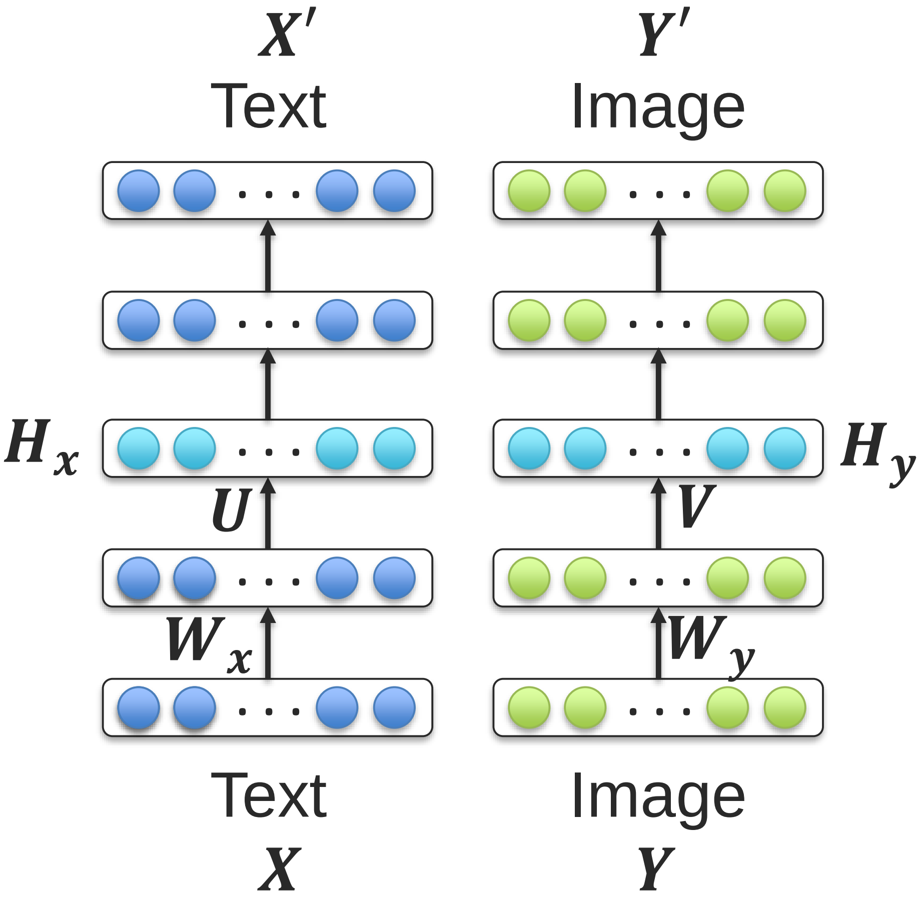


Andrew et al., ICML 2013

Deep Canonical Correlation Analysis

Training procedure:

1. Pre-train the models parameters using denoising autoencoders

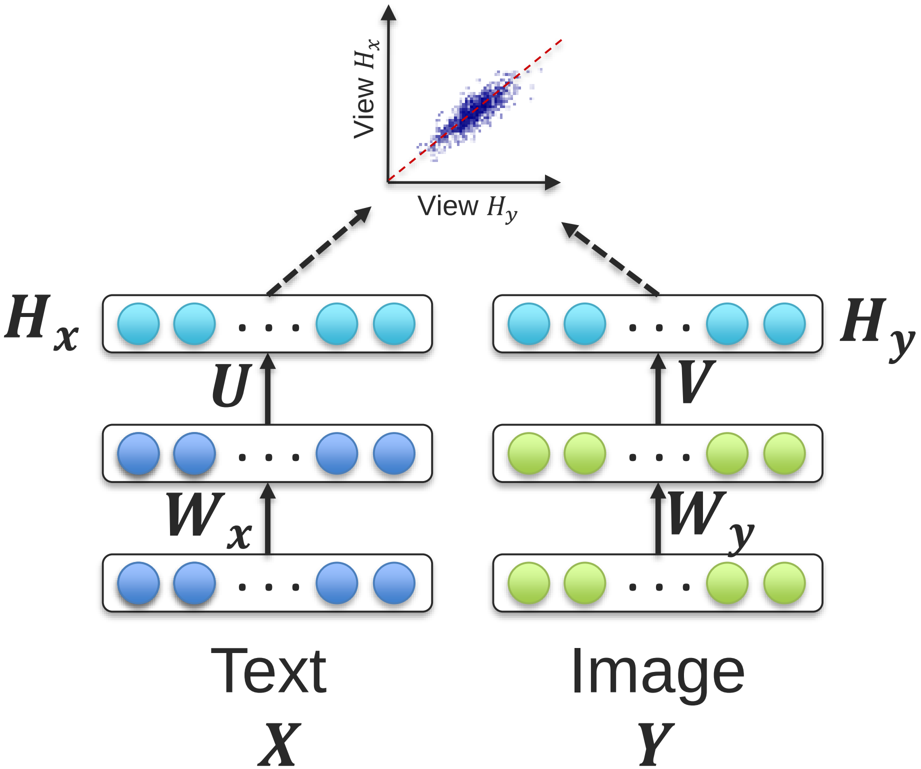


Andrew et al., ICML 2013

Deep Canonical Correlation Analysis

Training procedure:

1. Pre-train the models parameters using denoising autoencoders
2. Optimize the CCA objective functions using large mini-batches or full-batch (L-BFGS)

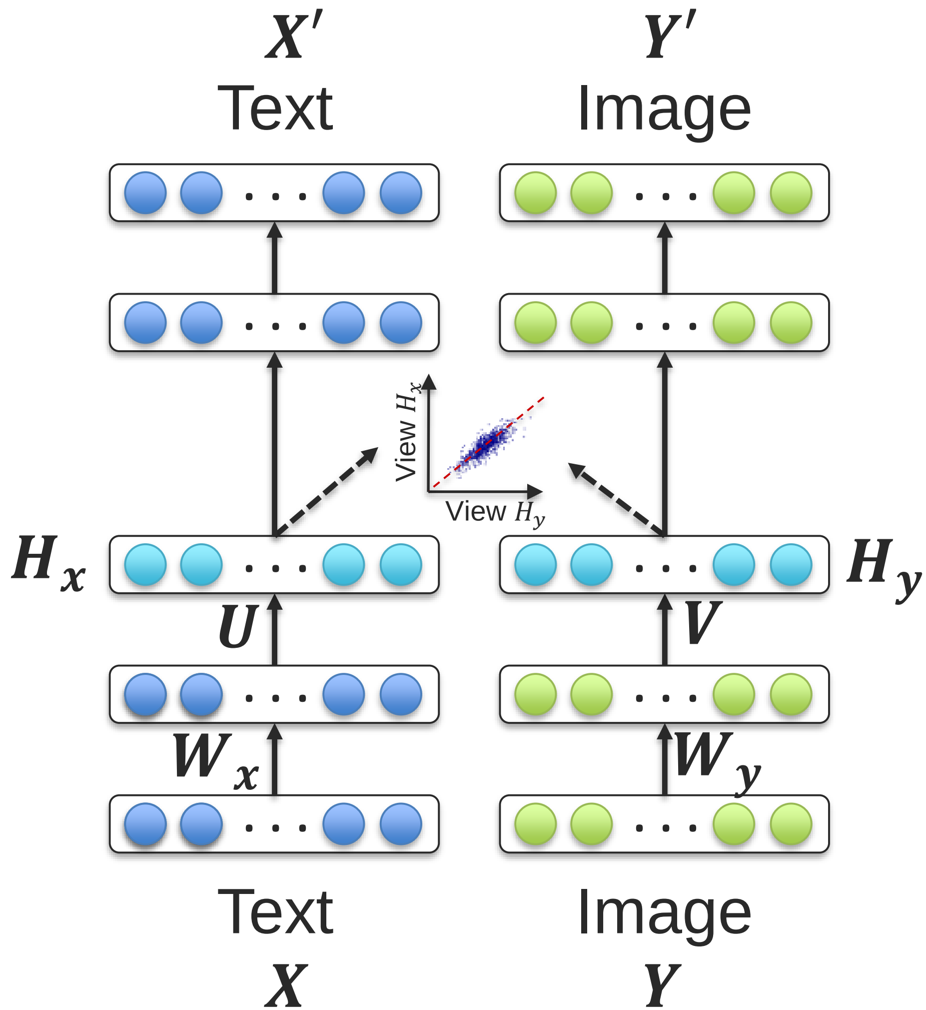


Andrew et al., ICML 2013

Deep Canonically Correlated Autoencoders (DCCAE)

Jointly optimize for DCCA and autoencoders loss functions

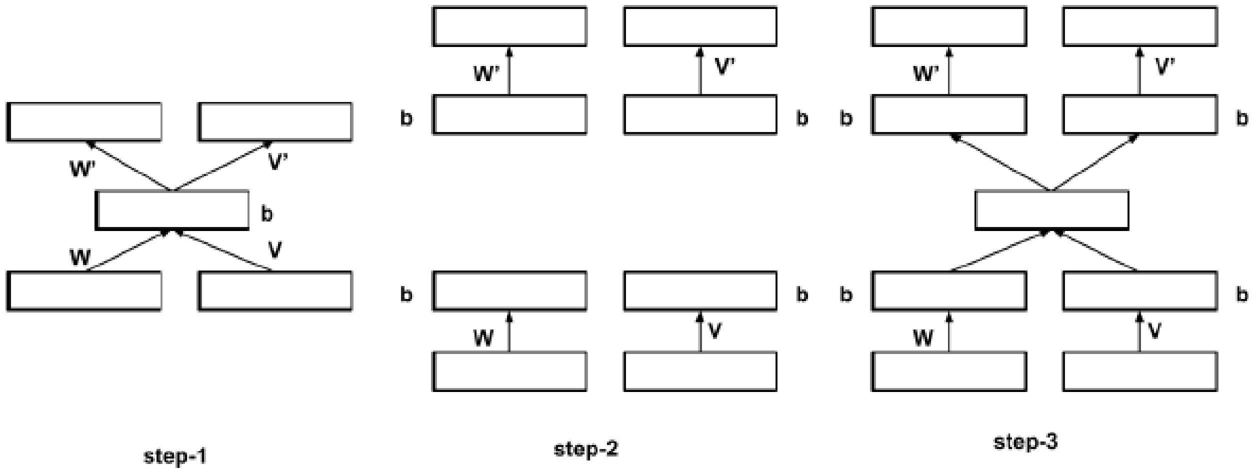
- A trade-off between multi-view correlation and reconstruction error from individual views



Wang et al., ICML 2015

Deep Correlational Neural Network

1. Learn a shallow CCA autoencoder (similar to 1 layer DCCAE model)
2. Use the learned weights for initializing the autoencoder layer
3. Repeat procedure



Chandar et al., Neural Computation, 2015

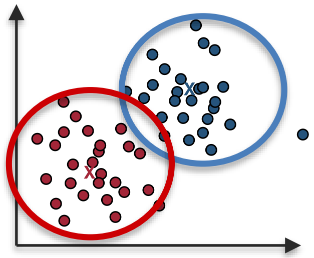
Matrix Factorization

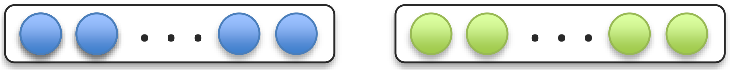
Data Clustering

How to discover groups in your data?

K-mean is a simple clustering algorithm based on competitive learning

- Iterative approach
  - Assign each data point to one cluster (based on distance metric)
  - Update cluster centers
  - Until convergence
- “Winner takes all”



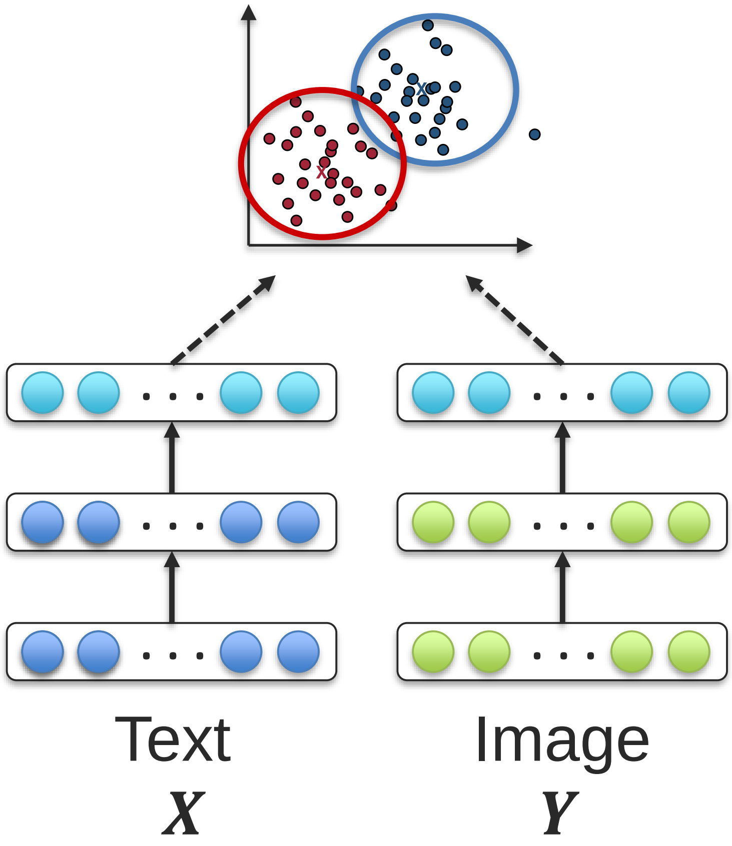


Text
X

Image
Y

Enforcing Data Clustering in Deep Networks

How to enforce data clustering in our (multimodal) deep learning algorithms?



Nonnegative Matrix Factorization (NMF)

Given: Nonnegative n x m matrix M (all entries ≥ 0)

$$\begin{pmatrix} \\ & X & \\ \\ \end{pmatrix} = \begin{bmatrix} \\ F \\ \\ \end{bmatrix} \begin{bmatrix} & G & \end{bmatrix}$$

Want: Nonnegative matrices F (n x r) and G (r x m), s.t. X = FG.

- easier to interpret
- provide better results in information retrieval, clustering

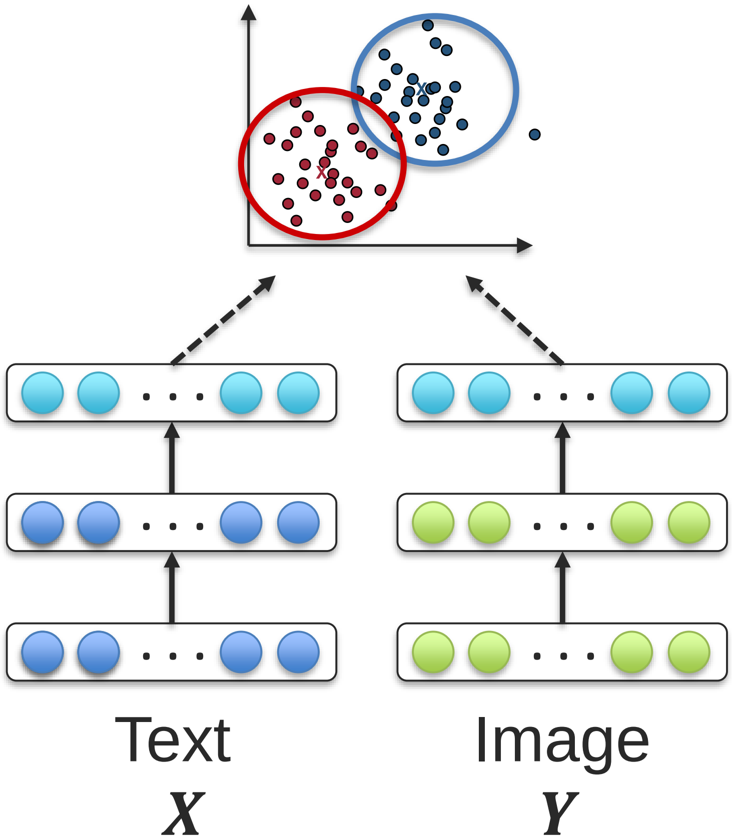
Semi-NMF and Other Extensions

SVD: $X_{\pm} \approx F_{\pm} G_{\pm}^T$

NMF: $X_{+} \approx F_{+} G_{+}^T$

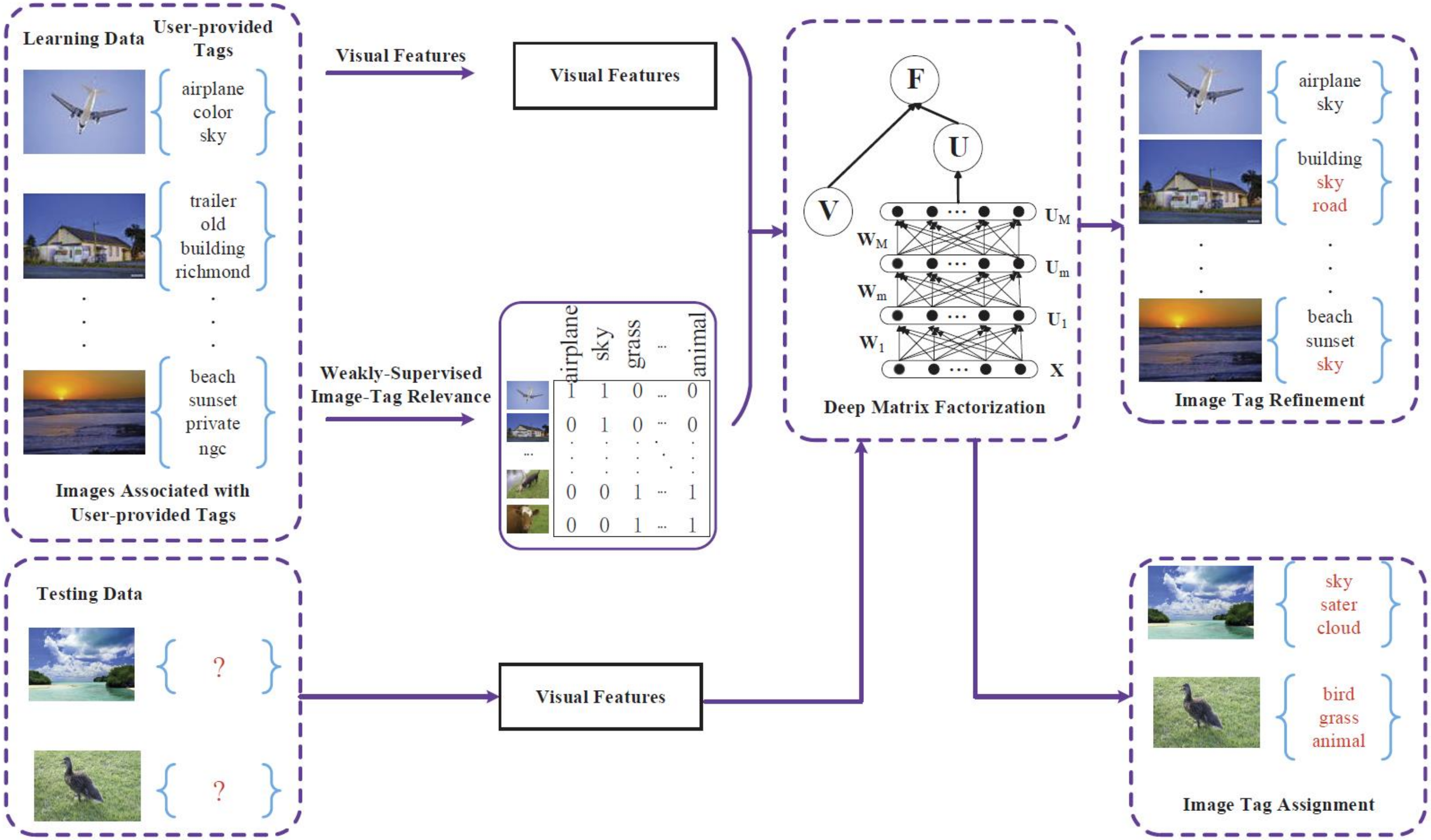
Semi-NMF: $X_{\pm} \approx F_{\pm} G_{+}^T$

Convex-NMF: $X_{\pm} \approx X_{\pm} W_{+} G_{+}^T$



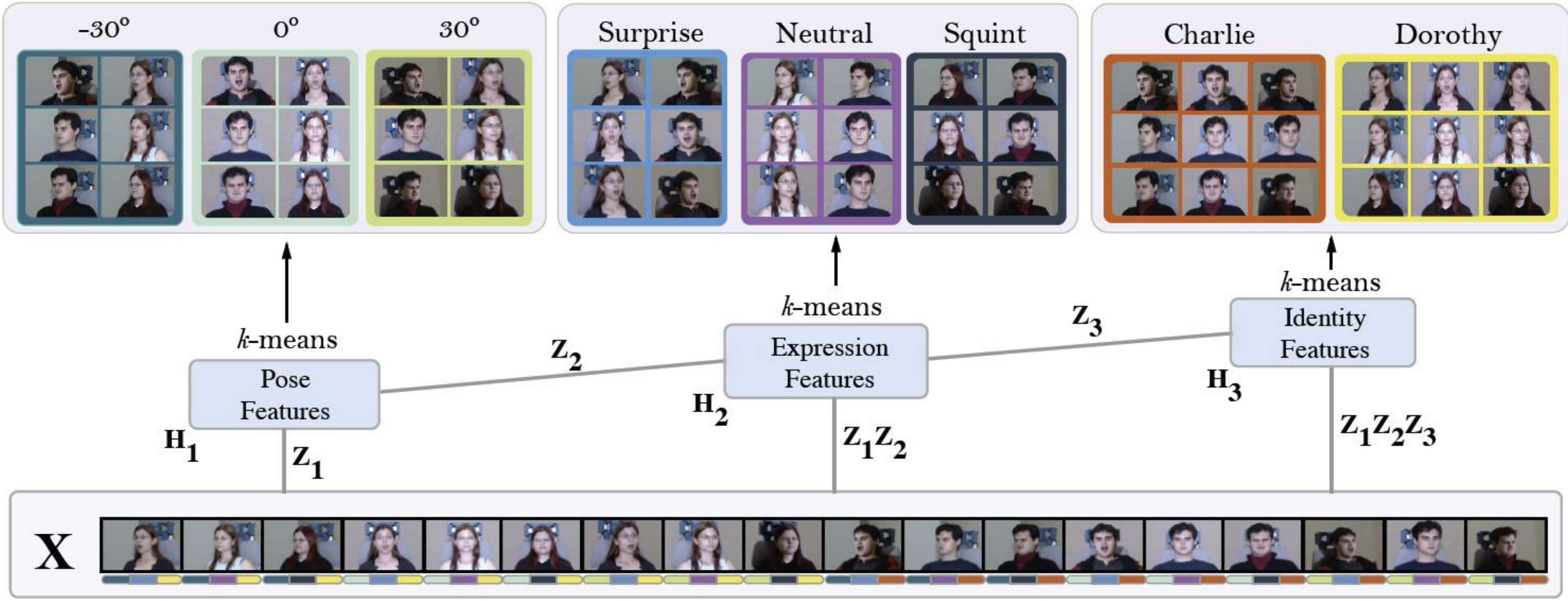
Ding et al., TPAMI2015

Deep Matrix Factorization



Li and Tang, MMML 2015

Deep Semi-NMF Model



Trigerous et al., TPAMI 2015

Multivariate Statistics

- Multivariate analysis of variance (MANOVA)
- Principal components analysis (PCA)
- Factor analysis
- Linear discriminant analysis (LDA)
- Canonical correlation analysis (CCA)
- Correspondence analysis
- Canonical correspondence analysis
- Multidimensional scaling
- Multivariate regression
- Discriminant analysis