Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

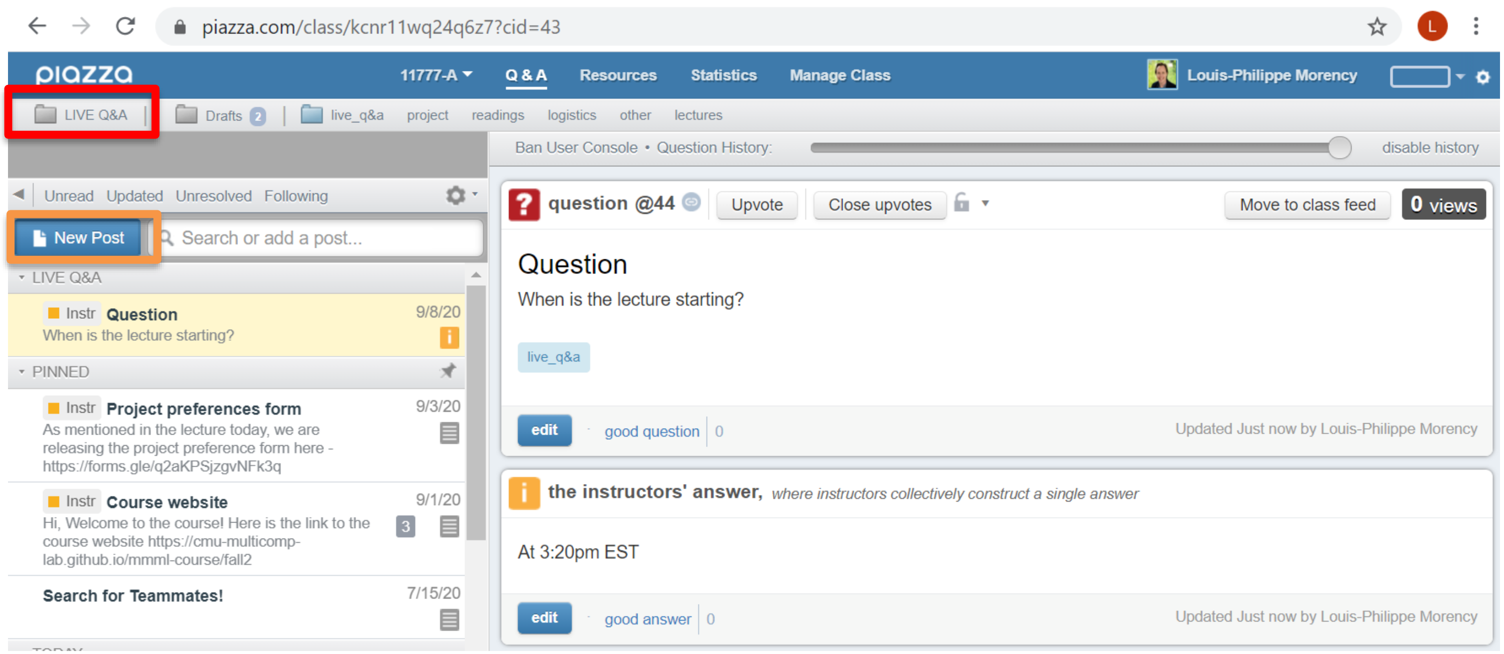
Lecture 4.2: Coordinated Representations

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Administrative Stuff

Piazza Live Q&A – Reminder



Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 1 9/1 & 9/3	**Course introduction** • Research and technical challenges • Course syllabus and requirements	**Multimodal applications and datasets** • Research tasks and datasets • Team projects
Week 2 9/8 & 9/10	**Basic concepts: neural networks** • Language, visual and acoustic • Loss functions and neural networks	**Basic concepts: network optimization** • Gradients and backpropagation • Practical deep model optimization
Week 3 9/15 & 9/17	**Visual unimodal representations** • Convolutional kernels and CNNs • Residual network and skip connection	**Language unimodal representations** • Gated networks and LSTM • Backpropagation Through Time
Week 4 9/22 & 9/24	**Multimodal representation learning** • Multimodal auto-encoders • Multimodal joint representations	**Coordinated representations** • Deep canonical correlation analysis • Non-negative matrix factorization
Week 5 9/29 & 10/1	**Multimodal alignment** • Explicit - dynamic time warping • Implicit - attention models	**Alignment and representation** • Self-attention models • Multimodal transformers
Week 6 10/6 & 10/8	***First project assignment*** *(live working sessions instead of lectures)*	

First project assignment
Presentations due Friday 10/9
Reports due Sunday 10/11
Peer feedback due Friday 10/16

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 7 10/13 & 10/15	**Alignment and translation** • Module networks • Connectionist temporal classification	**Probabilistic graphical models** • Dynamic Bayesian networks • Coupled and factor HMMs
Week 8 10/20 & 10/22	**Discriminative graphical models** • Conditional random fields • Continuous and fully-connected CRFs	**Neural Generative Models** • Variational auto-encoder • Generative adversarial networks
Week 9 10/27 & 10/29	**Reinforcement learning** • Markov decision process • Q learning and policy gradients	**Multimodal RL** • Deep Q learning • Multimodal applications
Week 10 11/3 & 11/5	**Fusion and co-learning** • Multi-kernel learning and fusion • Few shot learning and co-learning	**New research directions** • Recent approaches in multimodal ML
Week 11 11/10 & 11/12	***Mid-term project assignment*** *(live working sessions instead of lectures)*	

Midterm project assignment
Presentations due Friday 11/13
Reports due Sunday 11/15
Peer feedback due Friday 11/20

Lecture Schedule

Classes	Tuesday Lectures	Thursday Lectures
Week 12 11/17 & 11/19	**Embodied Language Grounding** • Connecting Language to Action • Guest lecture: Yonatan Bisk	**Multi-lingual representations** • Tentative topic • Guest lecture: To be confirmed
Week 13 11/24 & 11/26	***Thanksgiving week*** *(no lectures)*	
Week 14 12/1 & 12/3	**Learning to connect text and images** • Discourse approaches, text & images • Guest lecture: Malihe Alikhani	**Bias and fairness** • Computational ethics • Guest lecture: Yulia Tsvetkov
Week 15 12/8 & 12/10	***Final project assignment*** *(live working sessions instead of lectures)*	

Final project assignment
Presentations due Friday 12/11
Reports due Sunday 12/13

GPU $50 Coupons - AWS

- First, create an account on AWS Educate portal: https://aws.amazon.com/education/awseducate/
- Your account will need to be backed by your credit card

Be sure to setup billing alarms and monitor your spending!

- Refrain from including AWS credential in code/github
- To get your coupon, use your AndrewID and the URL posted on Piazza

GPU $50 Coupons - GCP

Google Cloud Platform

- Coupons can be redeemed at this address: https://console.cloud.google.com/education

Be sure to setup billing alarms and monitor your spending!

- Refrain from including GCP credential in code/github
- To get your coupon, use your AndrewID and the URL posted on Piazza

Multimodal Machine Learning

Lecture 4.2: Coordinated Representations

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

Lecture Objectives

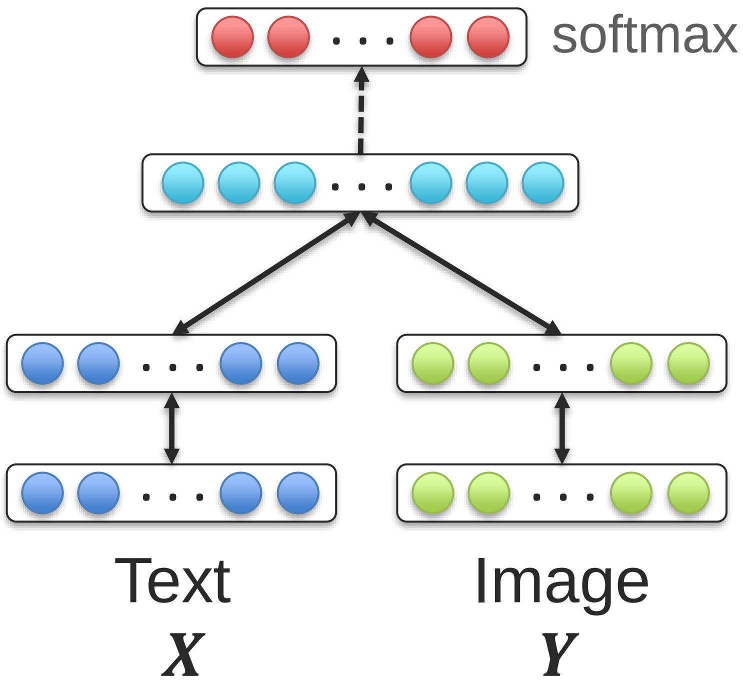
- Quick recap
- Coordinated multimodal representations
- Multivariate statistical analysis
  - Basic concepts (multivariate, covariance,…)
- Canonical Correlation Analysis
  - Deep Correlation Networks
  - Deep CCA, DCCA-AutoEncoder
- Multi-view clustering
  - Nonnegative Matrix Factorization
- Multi-view latent intact space
  - Autoencoder in Autoencoder networkds

Quick Recap

Multimodal Representation Learning

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

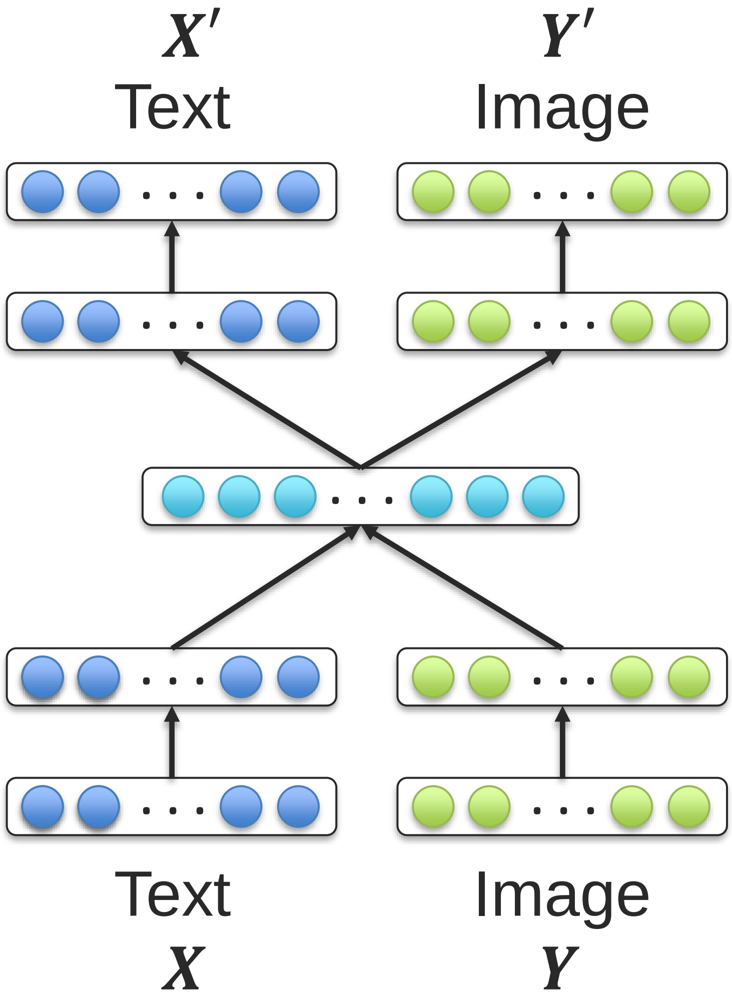
- ❑ Deep Multimodal Boltzmann machines



Multimodal Representation Learning

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

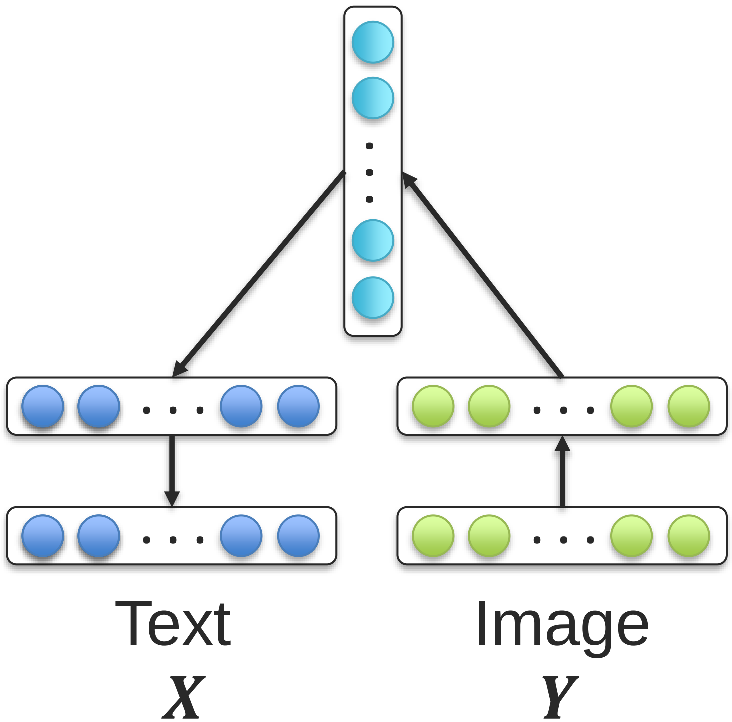
- ❑ Deep Multimodal Boltzmann machines
- ❑ Stacked Autoencoder



Multimodal Representation Learning

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

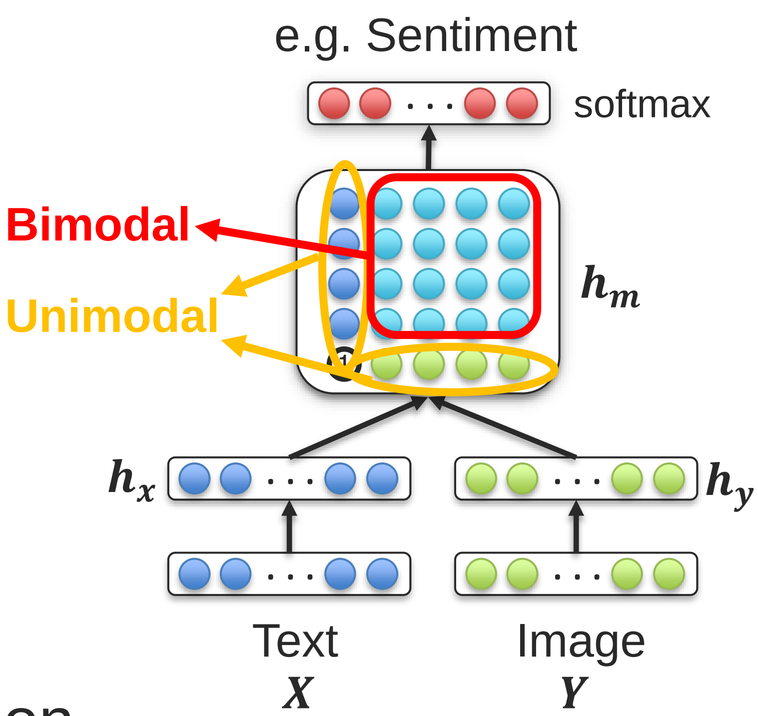
- ❑ Deep Multimodal Boltzmann machines
- ❑ Stacked Autoencoder
- ❑ Encoder-Decoder



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

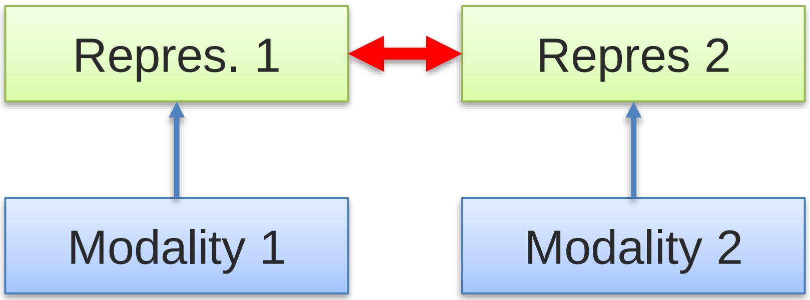


How Can We Learn Better Representations?

Coordinated Multimodal Representations

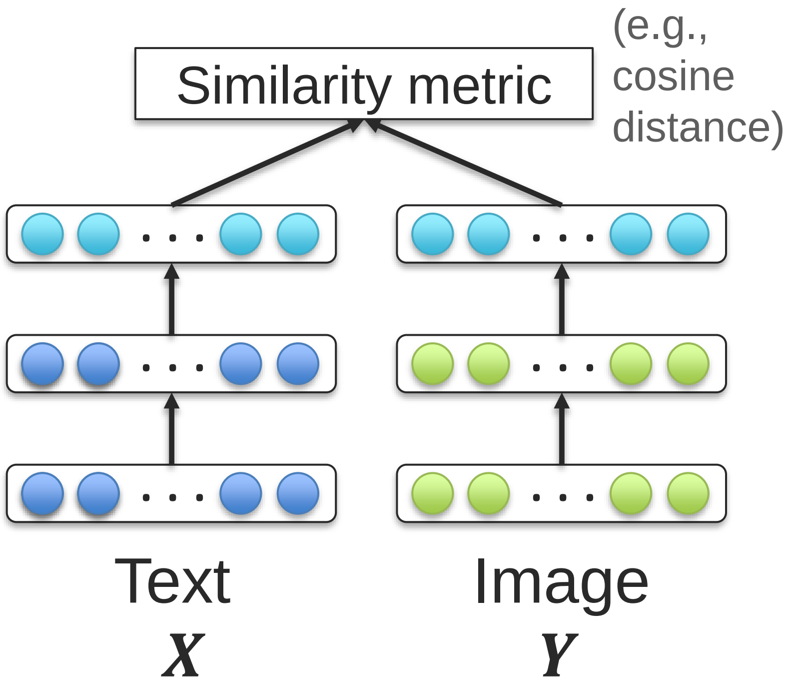
Coordinated multimodal embeddings

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

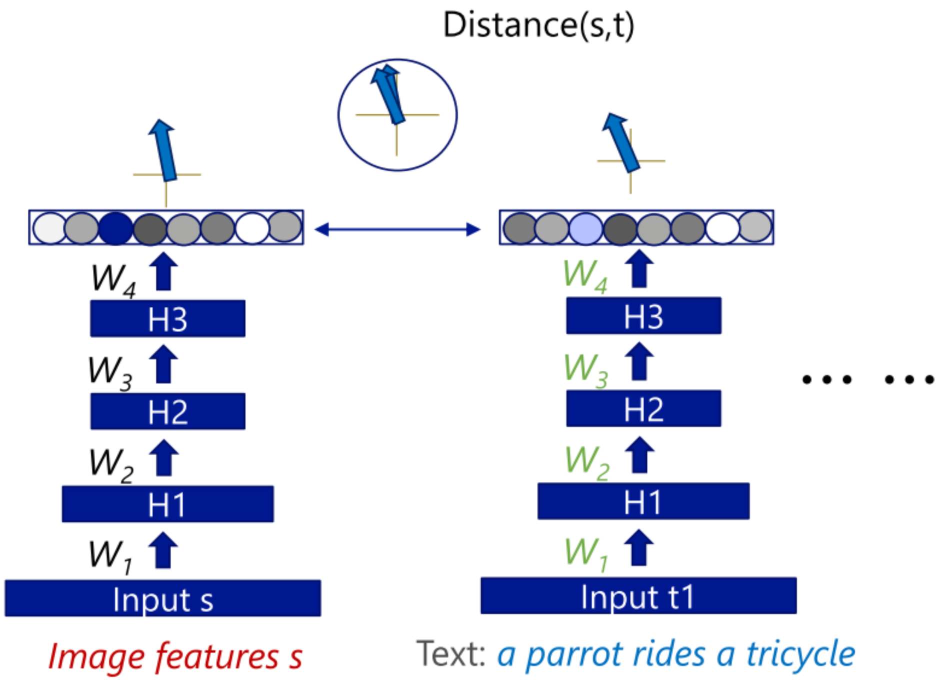


Coordinated Multimodal Representations

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

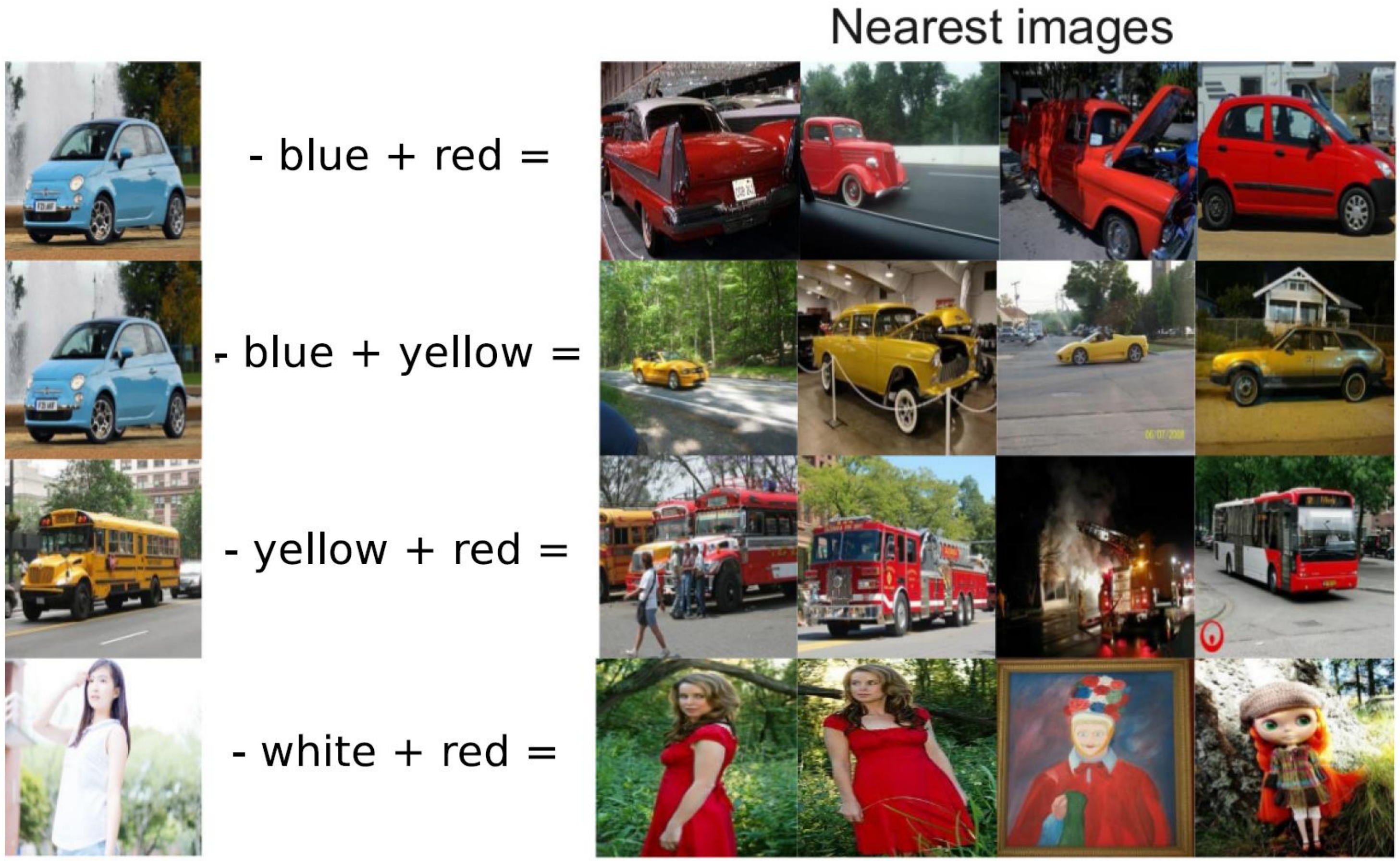


Coordinated Multimodal Embeddings



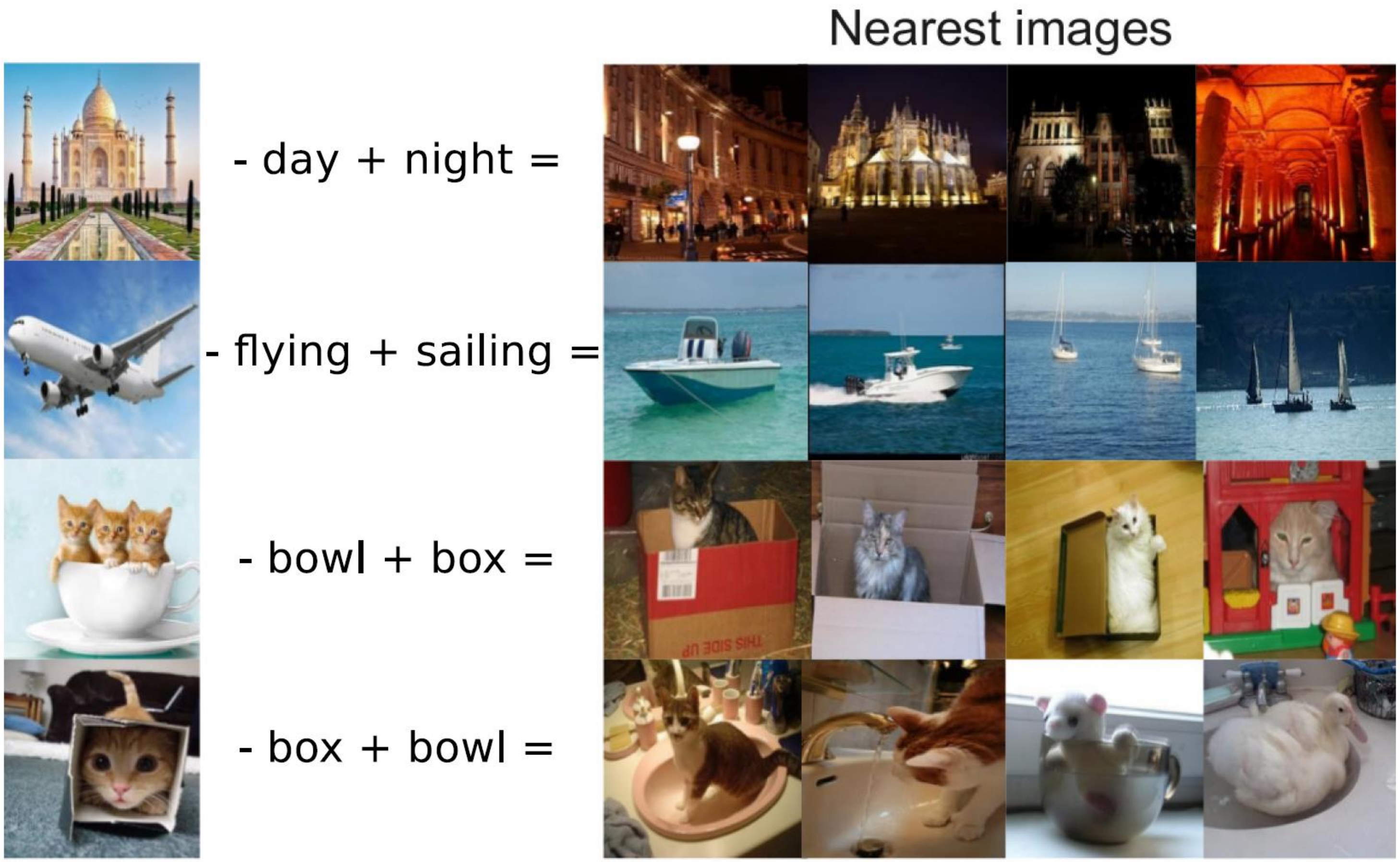
[Huang et al., Learning Deep Structured Semantic Models for Web Search using Clickthrough Data, 2013]

Multimodal Vector Space Arithmetic



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

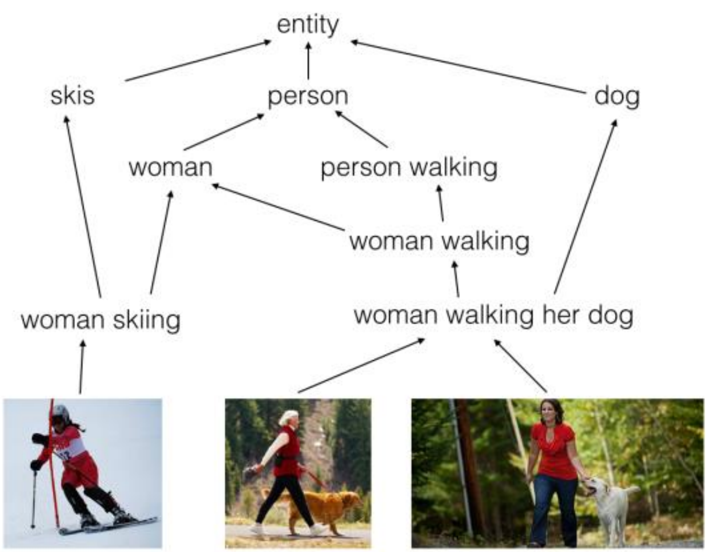
Multimodal Vector Space Arithmetic

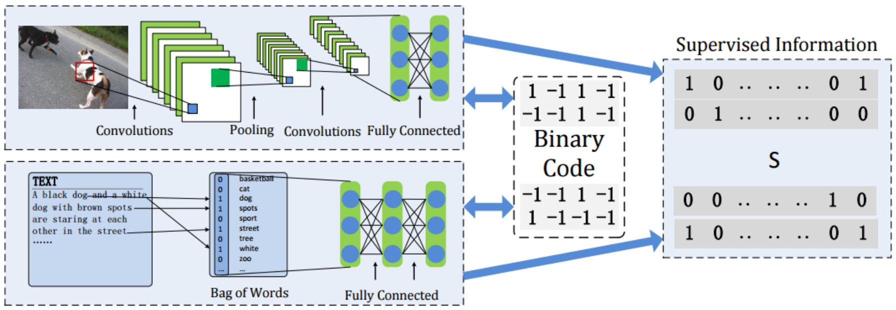


[Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, 2014]

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

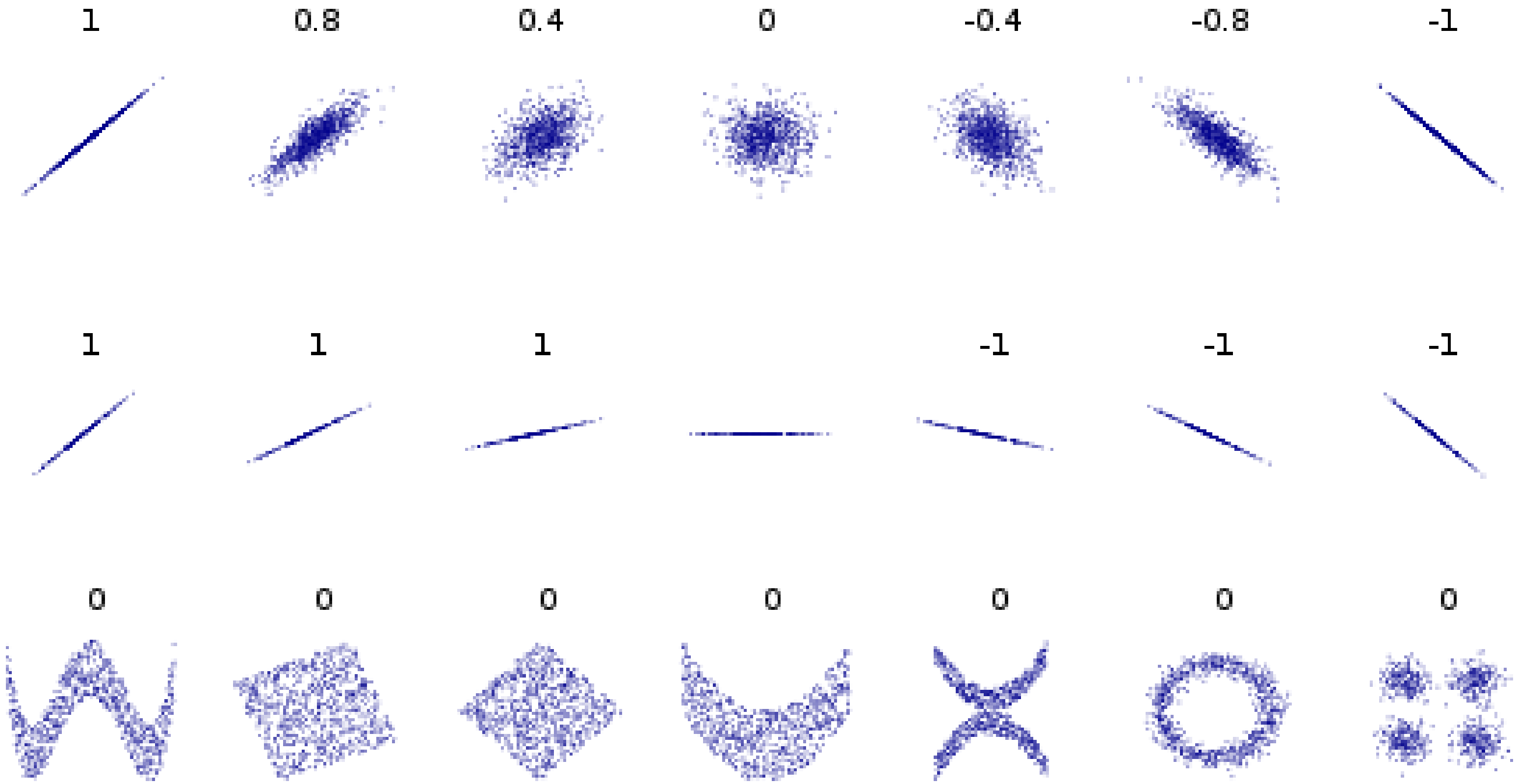
$$cov(X,Y) = E[(X-\mu_X)(Y-\mu_y)] = E[\bar{X}\bar{Y}]$$

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

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

Cross-covariance matrix generalizes the notion of covariance

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

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

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

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

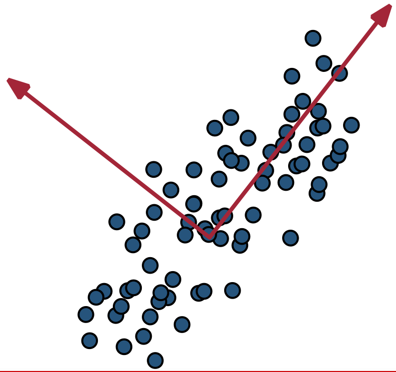
Trace is defined as the sum of the elements on the main diagonal of any matrix $\boldsymbol{X}$

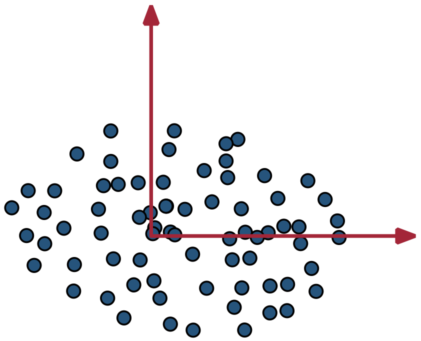
$$tr(\boldsymbol{X}) = \sum_{i=1}^{n} x_{ii}$$

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 **x** in R^n such that $A\mathbf{x}$ is a scalar multiple of **x**?

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

Eigenvalue equation:

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

Eigenvector

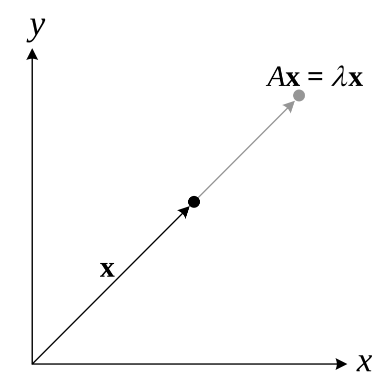
Eigenvalue

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 $\mathbf{A}^T\mathbf{A}$.

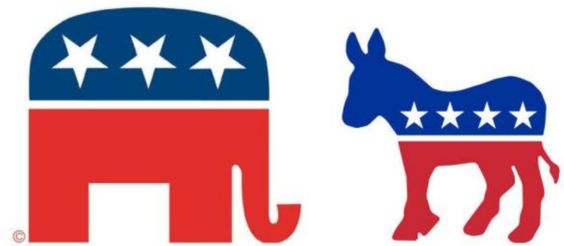
$$\mathbf{A}\mathbf{A}^T\mathbf{u}_i = s_i^2\mathbf{u}_i$$
$$\mathbf{A}^T\mathbf{A}\mathbf{v}_i = s_i^2\mathbf{v}_i$$

Canonical Correlation Analysis

Multi-view Learning

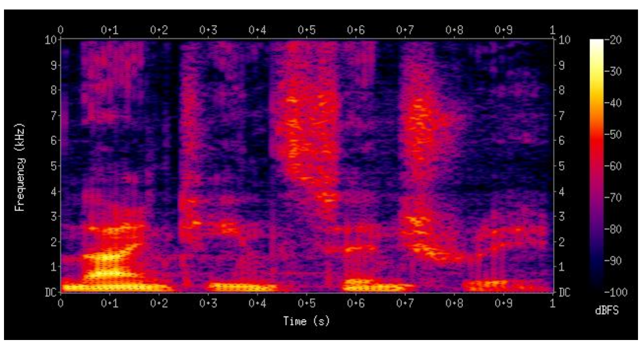
X

Y



demographic properties

responses to survey



audio features at time i

video features at time i

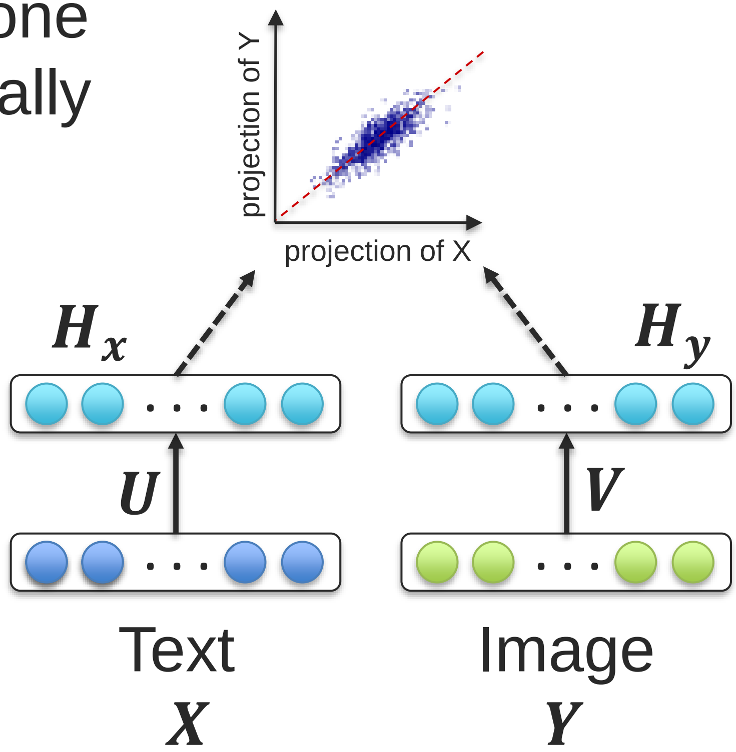
Canonical Correlation Analysis

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

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

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

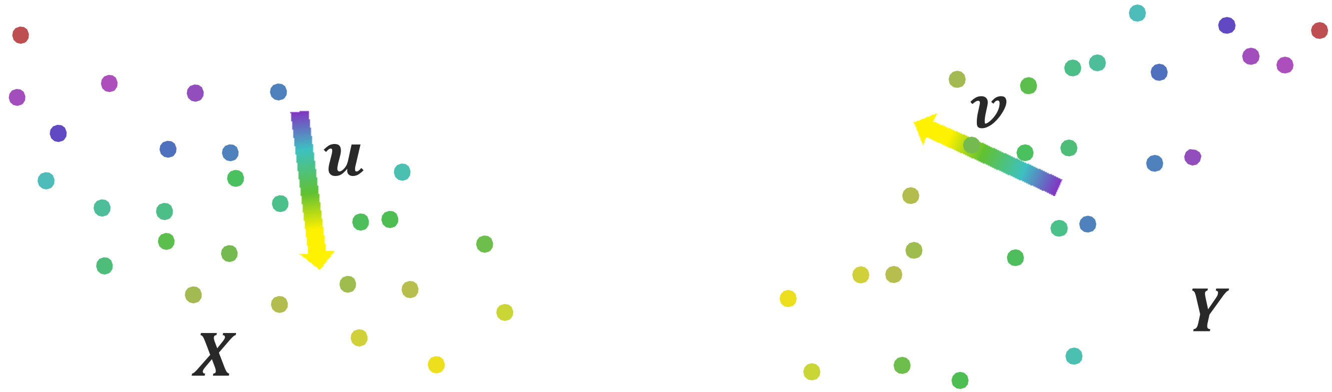
$$= \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr(\boldsymbol{u^T X}, \boldsymbol{v^T Y})$$



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

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 $(\boldsymbol{u}_{(i)}\boldsymbol{X}, \boldsymbol{v}_{(i)}\boldsymbol{Y})$:

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

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

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

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

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

Canonical Correlation Analysis

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

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

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

Canonical Correlation Analysis:

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

subject to: $\boldsymbol{U}^T \boldsymbol{\Sigma}_{XX} \boldsymbol{U} = \boldsymbol{V}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{V} = \boldsymbol{I},\ \boldsymbol{u}_{(j)}^T \boldsymbol{\Sigma}_{XY} \boldsymbol{v}_{(i)} = \boldsymbol{0}$ for $i \neq j$

Canonical Correlation Analysis

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

subject to: $\boldsymbol{U}^T \boldsymbol{\Sigma}_{XX} \boldsymbol{U} = \boldsymbol{V}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{V} = \boldsymbol{I}, \boldsymbol{u}_{(j)}^T \boldsymbol{\Sigma}_{XY} \boldsymbol{v}_{(i)} = \boldsymbol{0}$ for $i \neq j$

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

Canonical Correlation Analysis

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

subject to: $\boldsymbol{U}^T \boldsymbol{\Sigma}_{XX} \boldsymbol{U} = \boldsymbol{V}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{V} = \boldsymbol{I}, \; \boldsymbol{u}_{(j)}^T \boldsymbol{\Sigma}_{XY} \boldsymbol{v}_{(i)} = \boldsymbol{0}$ for $i \neq j$

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: $\boldsymbol{U}^T \boldsymbol{\Sigma}_{XX} \boldsymbol{U} = \boldsymbol{V}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{V} = \boldsymbol{I}, \boldsymbol{u}_{(j)}^T \boldsymbol{\Sigma}_{XY} \boldsymbol{v}_{(i)} = \boldsymbol{0}$ for $i \neq j$

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

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

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

Eigenvalues

Eigenvectors

Eigenvalue equations

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

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

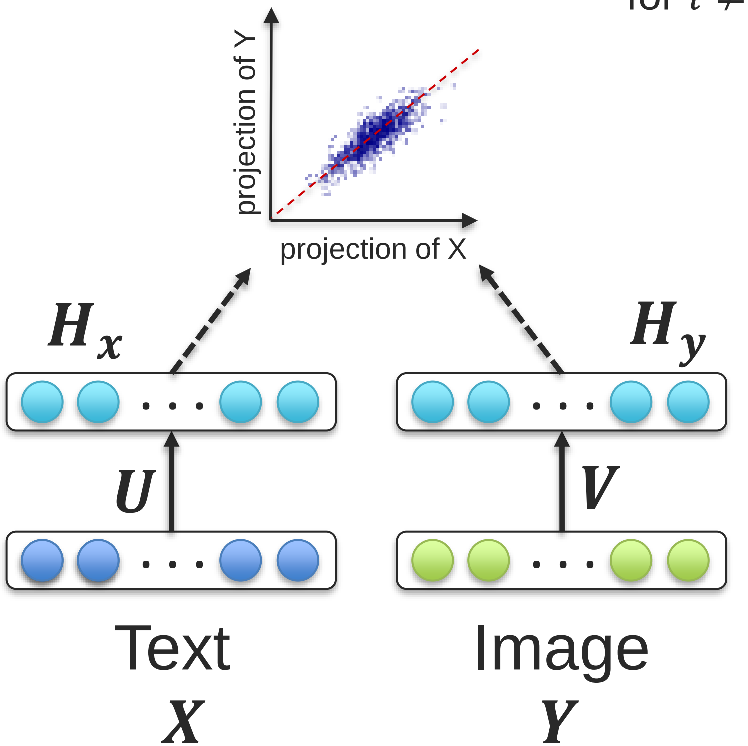
where $\lambda = 4\alpha\beta$

Canonical Correlation Analysis

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

subject to: $\boldsymbol{U}^T \boldsymbol{\Sigma}_{XX} \boldsymbol{U} = \boldsymbol{V}^T \boldsymbol{\Sigma}_{YY} \boldsymbol{V} = \boldsymbol{I},\ \boldsymbol{u}_{(j)}^T \boldsymbol{\Sigma}_{XY} \boldsymbol{v}_{(i)} = \boldsymbol{0}$ for $i \neq j$

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



Exploring Deep Correlation Networks

Deep Canonical Correlation Analysis

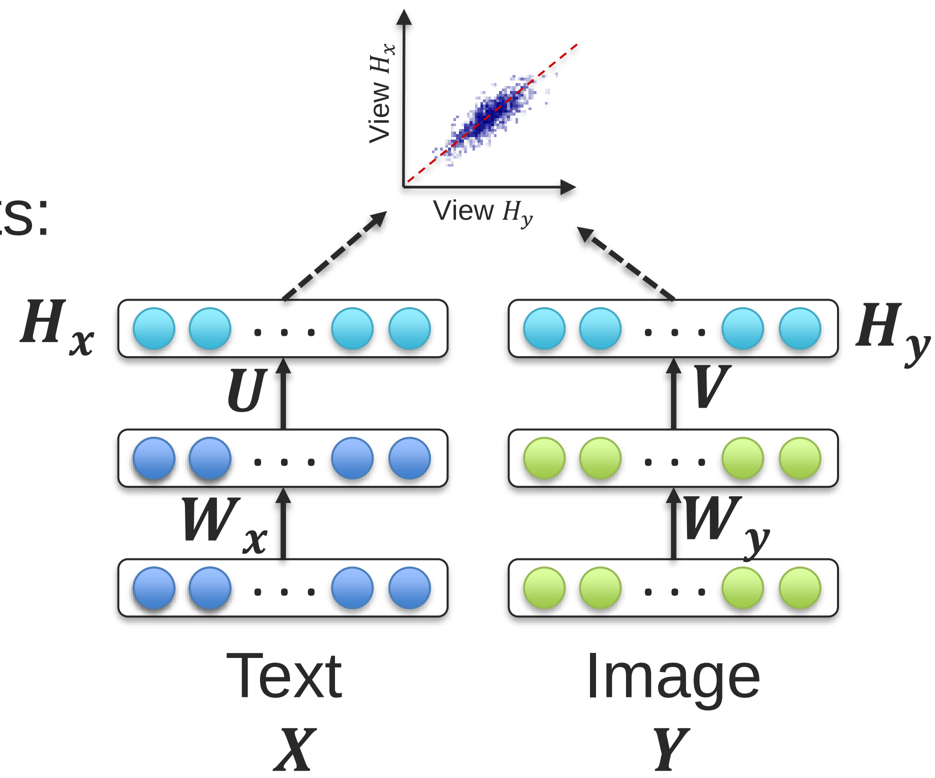
Same objective function as CCA:

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

And need to compute gradients:

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

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

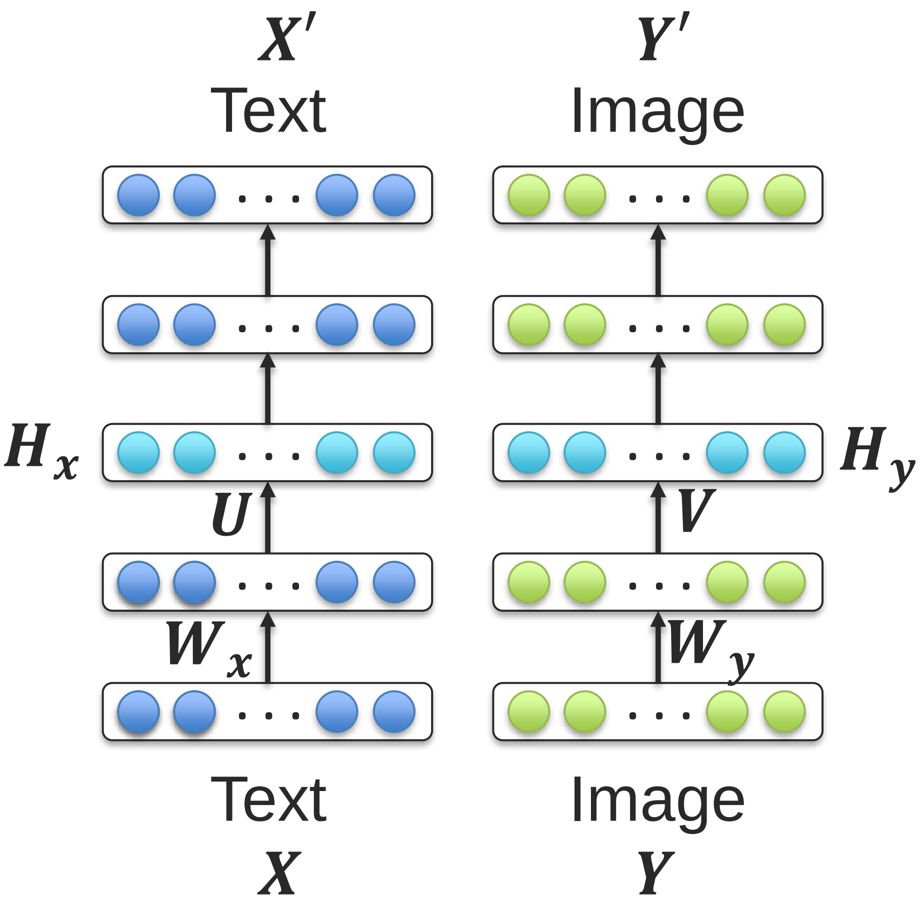


Andrew et al., ICML 2013

Deep Canonical Correlation Analysis

Training procedure:

1. Pre-train the models parameters using denoising autoencoders

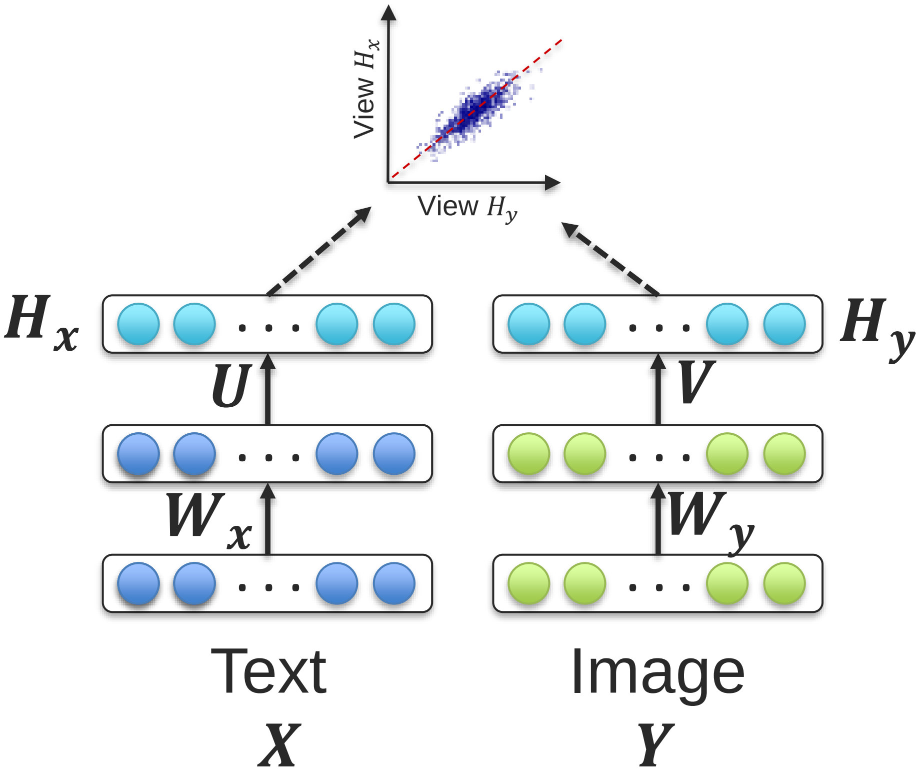


Andrew et al., ICML 2013

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)

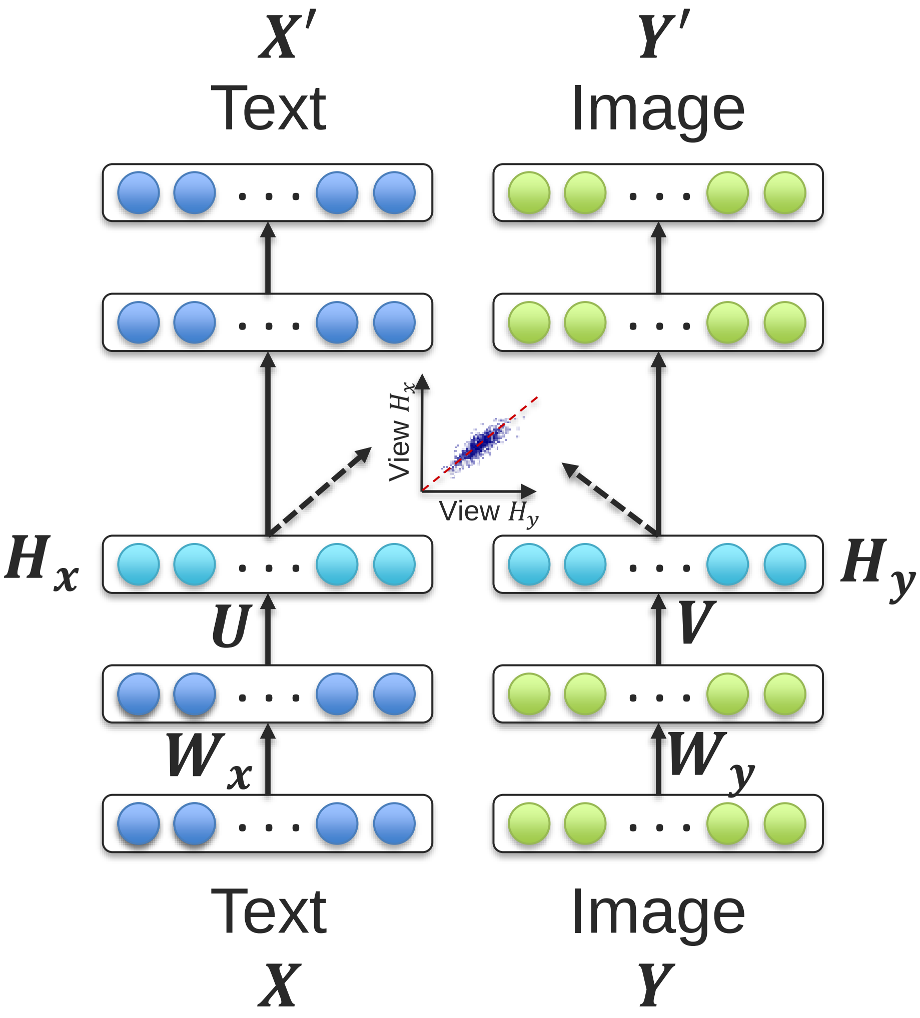


Andrew et al., ICML 2013

Deep Canonically Correlated Autoencoders (DCCAE)

Jointly optimize for DCCA and autoencoders loss functions

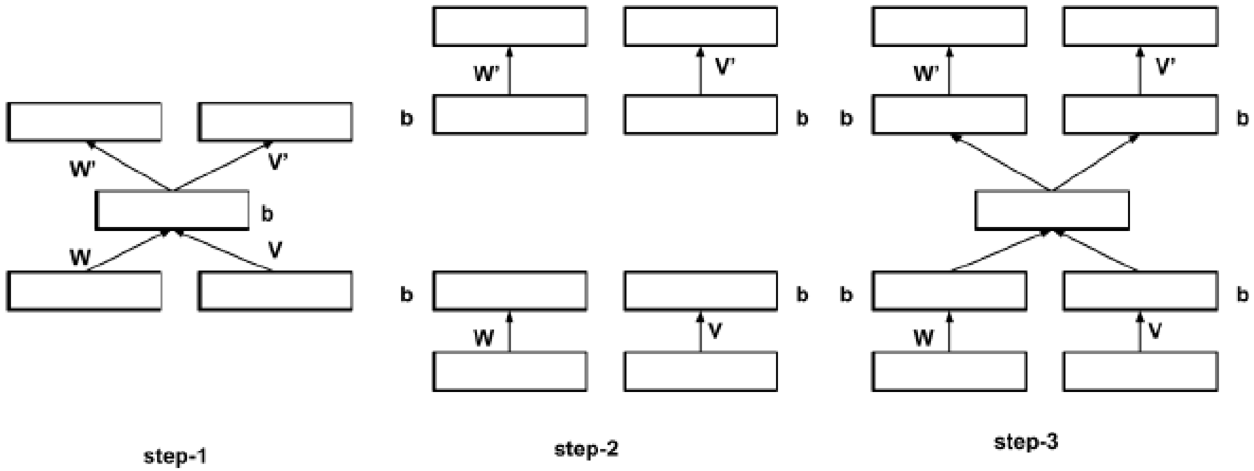
- A trade-off between multi-view correlation and reconstruction error from individual views



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

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

Multi-View Clustering

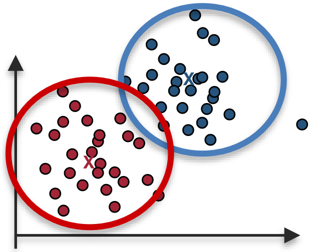
Data Clustering

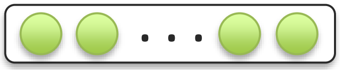
Clustering definition: partition a set of data samples such that similar samples are grouped, and dissimilar samples are divided

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”





Image

“Soft” Clustering: Nonnegative Matrix Factorization

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

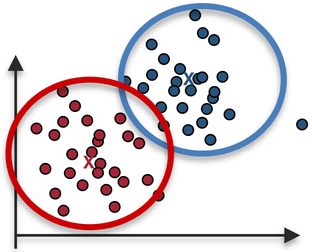
Semi-NMF and Other Extensions

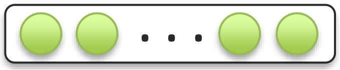
SVD: $X_{\pm} \approx F_{\pm} G_{\pm}^T$

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

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

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

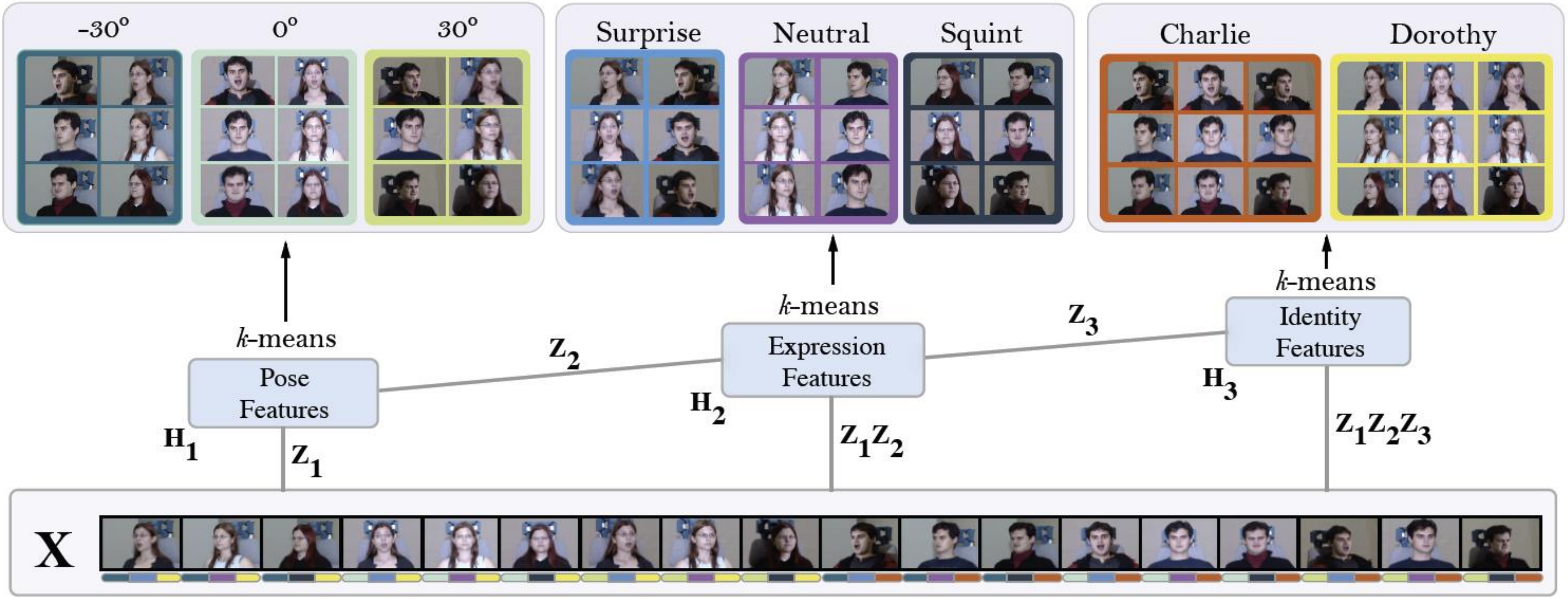




Image

Ding et al., TPAMI2015

Deep Semi-NMF Model



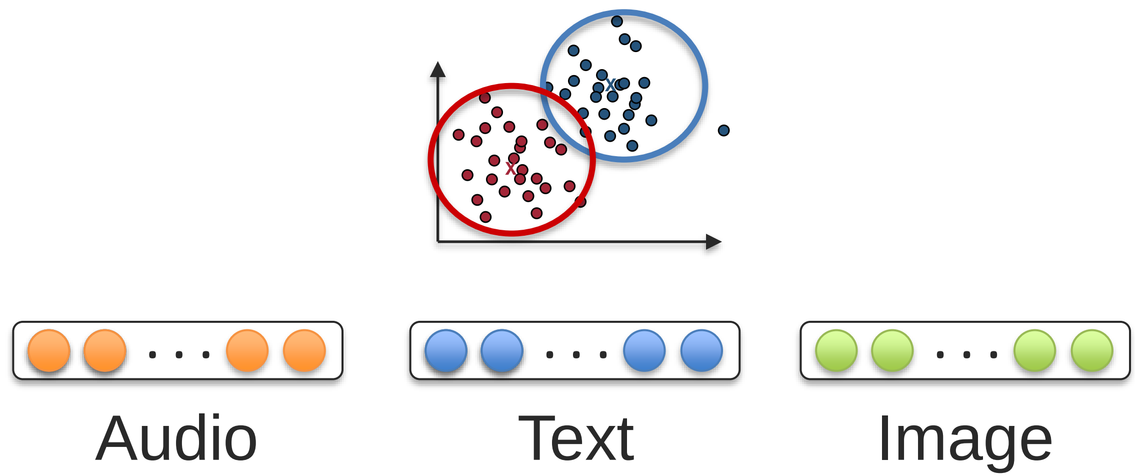
Trigerous et al., TPAMI 2015

Multi-View Clustering

Learn data partitioning from multiple views (modalities)

Views: different sources in diverse domains or obtained from various feature collectors or modalities

Example: Multiple views in computer vision - LBP, SIFT, HOG

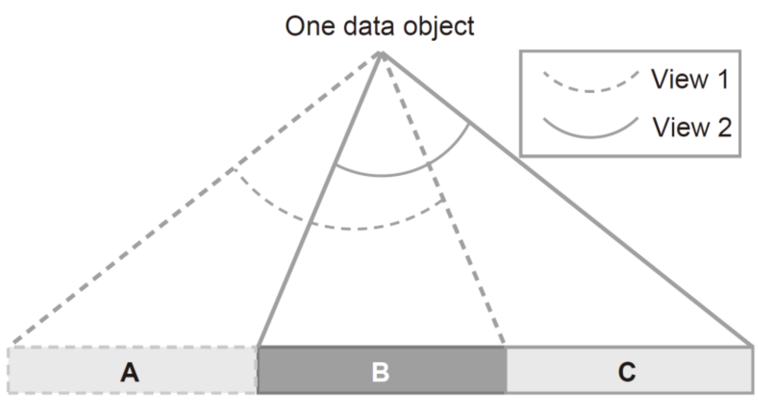


Yan Yang and Hao Wang, Multi-view Clustering: A Survey, Big data mining and analytics, Volume 1, Number 2, June 2018

Principles of Multi-View Clustering

Two important principles:

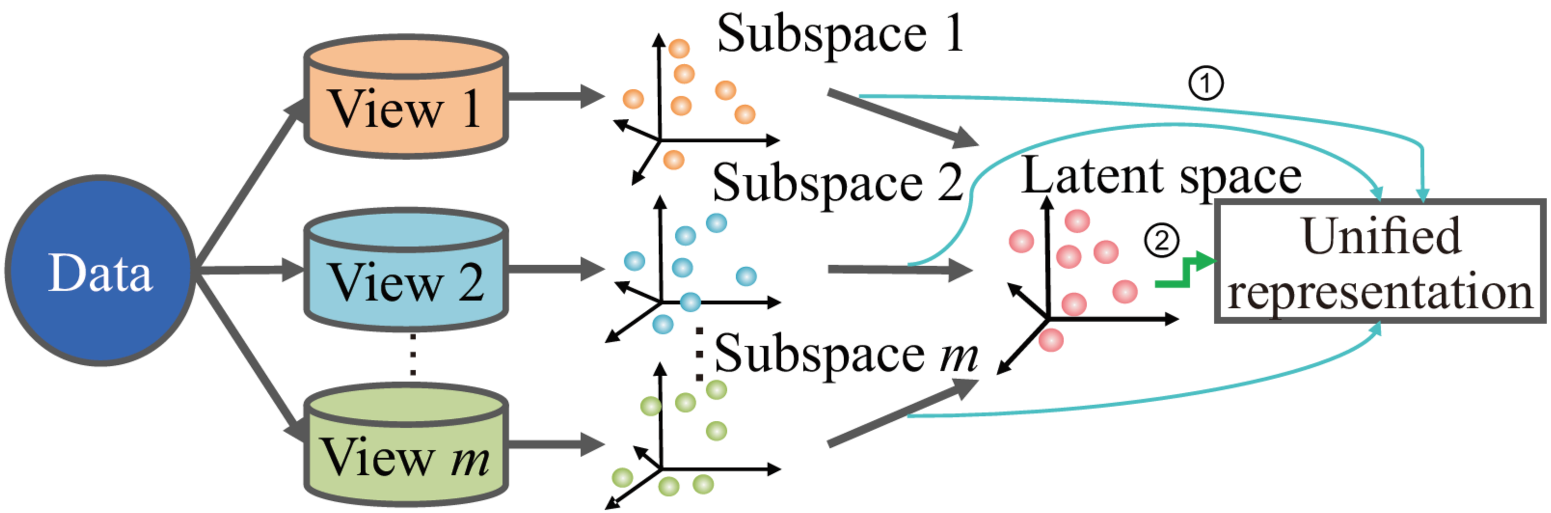
1. **Consensus principle:** maximize consistency across multiple distinct views
2. **Complementarity principle:** multiple views needed to get more comprehensive and accurate descriptions



Yan Yang and Hao Wang, Multi-view Clustering: A Survey, Big data mining and analytics, Volume 1, Number 2, June 2018

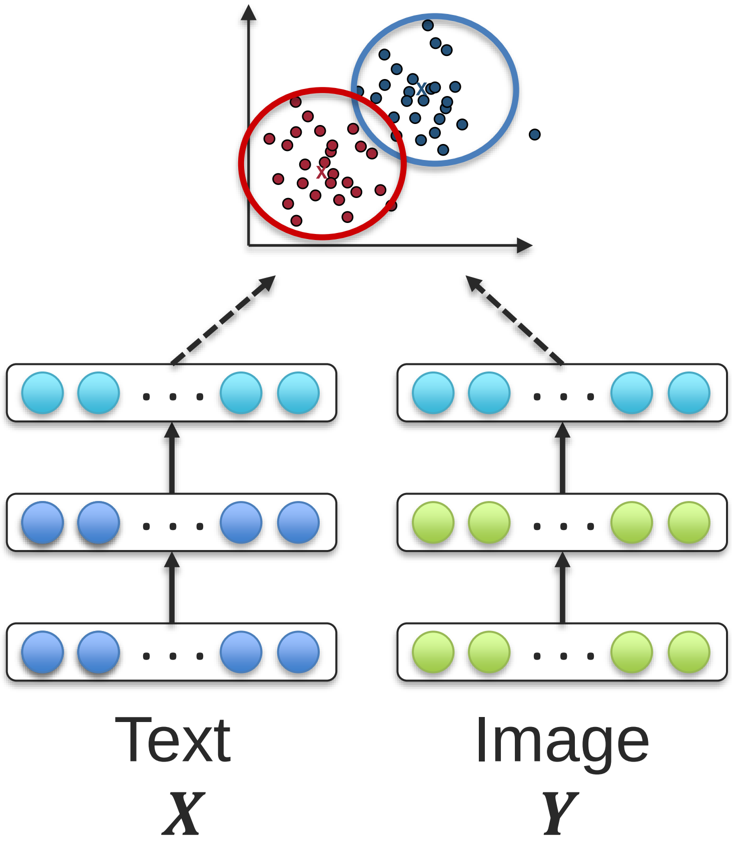
Multi-view subspace clustering

Definition: learns a unified feature representation from all the view subspaces by assuming that all views share this representation

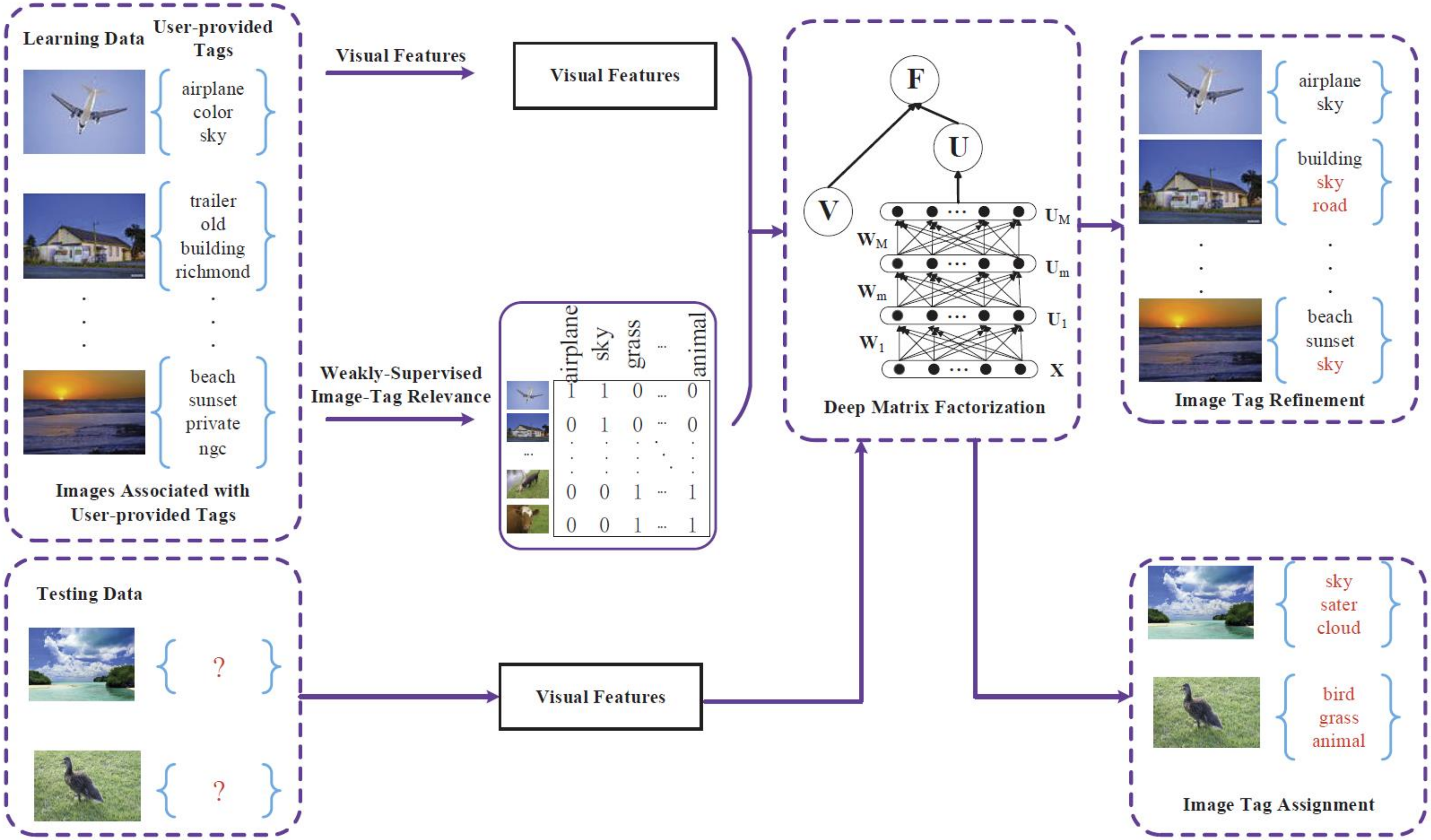


Enforcing Data Clustering in Deep Networks

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



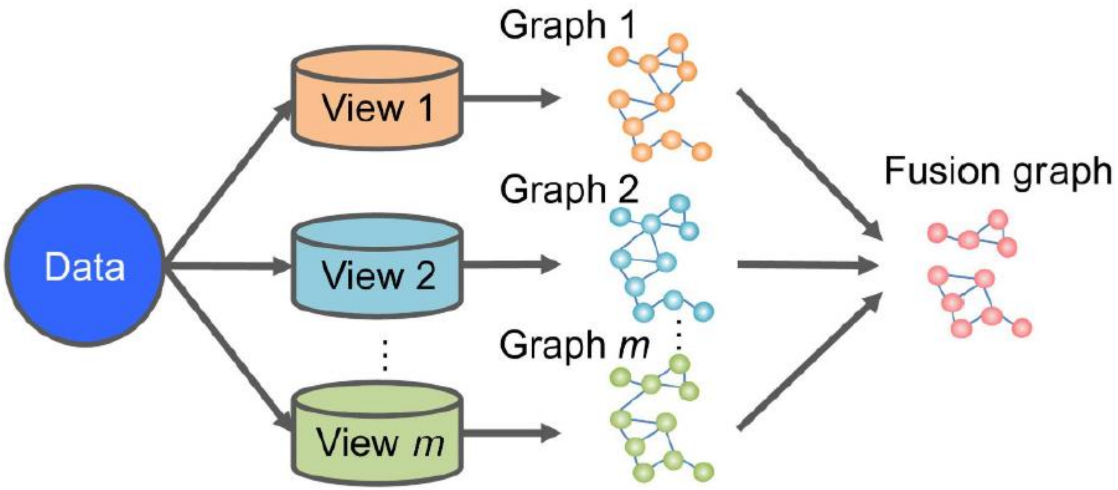
Deep Matrix Factorization



Li and Tang, MMML 2015

Other Multi-View Clustering Approaches

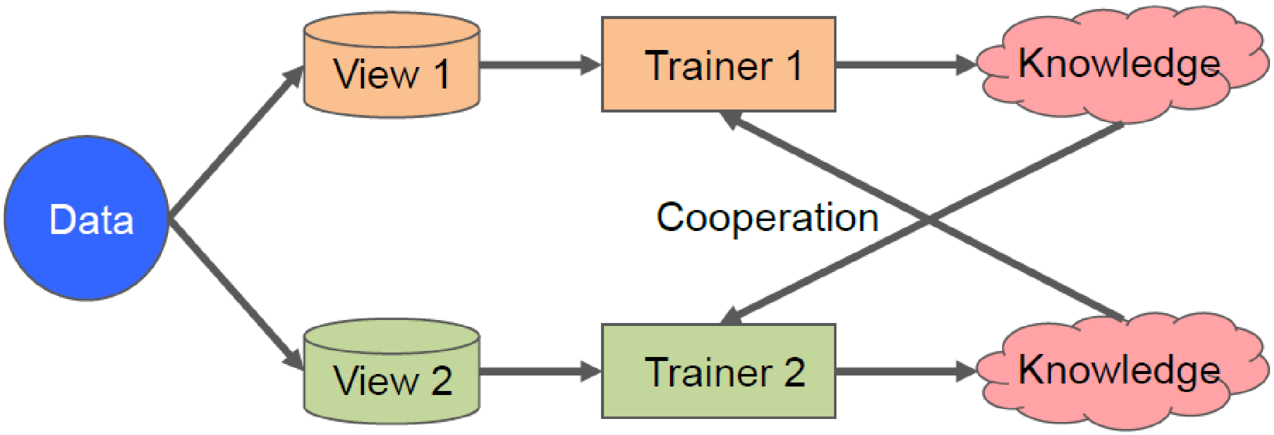
Graph-based clustering: search for a fusion graph (or network) across all views and then perform clustering



Yan Yang and Hao Wang, Multi-view Clustering: A Survey, Big data mining and analytics, Volume 1, Number 2, June 2018

Other Multi-View Clustering Approaches

Co-training: bootstraps the clustering of different views by using the learning knowledge from other views

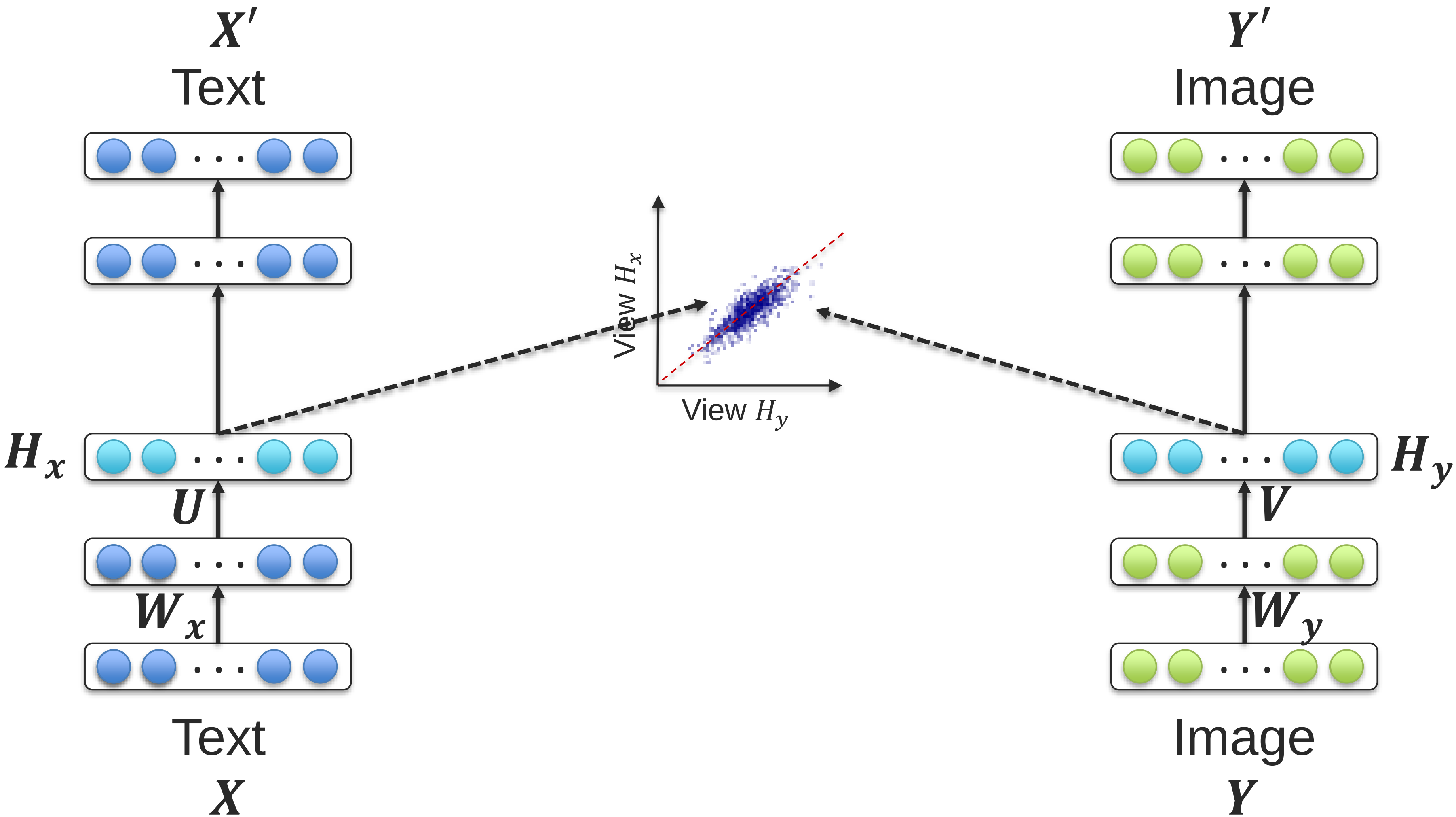


Yan Yang and Hao Wang, Multi-view Clustering: A Survey, Big data mining and analytics, Volume 1, Number 2, June 2018

Auto-Encoder in Auto-Encoder Network

Deep Canonically Correlated Autoencoders (DCCAE)

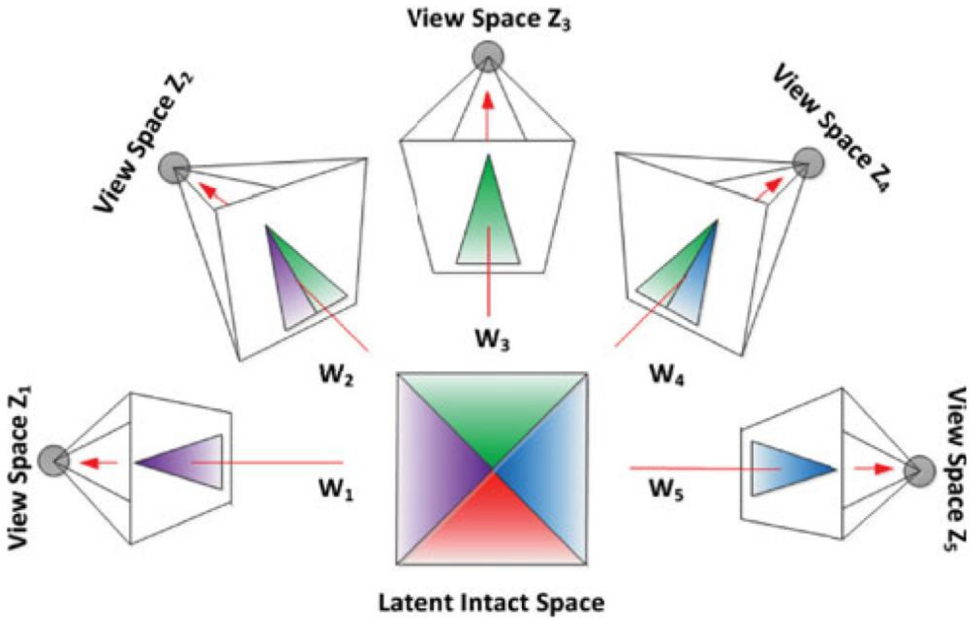
Wang et al., ICML 2015



Multi-view Latent "Intact" Space

Xu et al., TPAMI 2015

Given multiple views z_i from the same "object":

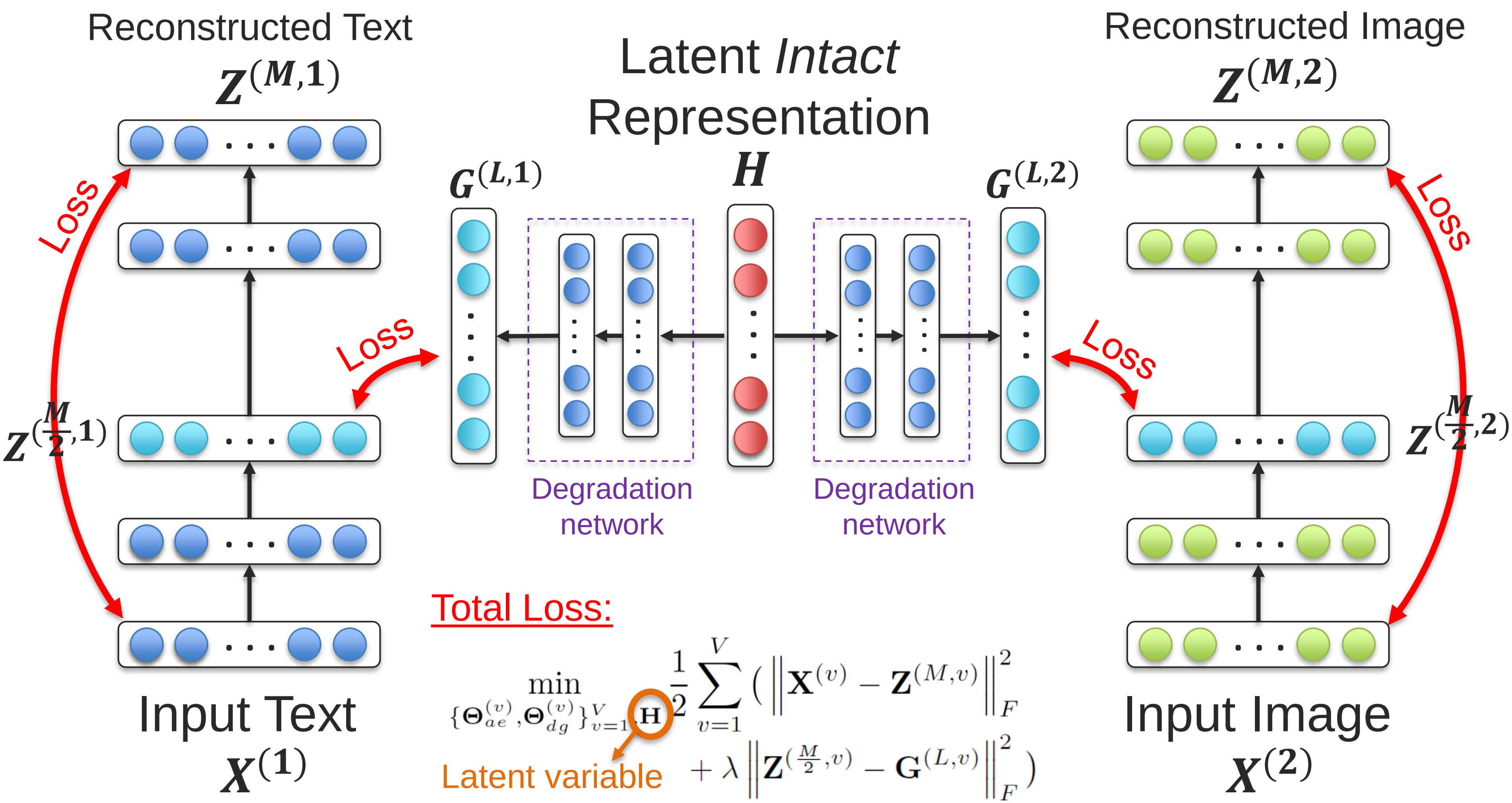


1) There is an "intact" representation which is *complete* and *not damaged*

2) The views z_i are partial (and possibly degenerated) representations of the intact representation

Auto-Encoder in Auto-Encoder Network

Zhang et al., CVPR 2019



Total Loss:

$$\min_{\{\Theta_{ae}^{(v)},\Theta_{dg}^{(v)}\}_{v=1}^{V},\mathbf{H}} \frac{1}{2}\sum_{v=1}^{V}\left(\left\|\mathbf{X}^{(v)}-\mathbf{Z}^{(M,v)}\right\|_F^2 + \lambda\left\|\mathbf{Z}^{(\frac{M}{2},v)}-\mathbf{G}^{(L,v)}\right\|_F^2\right)$$

Latent variable