Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.2: Multimodal Representations

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.*
Spring 2021 edition taught by Yonatan Bisk

Administrative Stuff

Upcoming Deadlines

- Today: Lecture highlight form
- Sunday: First project assignment
- Friday 10/1: Reading assignment
- Sunday 10/10: Second project assignment

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, contact your Primary TA on Piazza

GPU $50 Coupons - GCP

- 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, contact your Primary TA on Piazza

Language Technologies Institute

Carnegie Mellon University

Multimodal Machine Learning

Lecture 4.2: Multimodal Representations

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis.*
Spring 2021 edition taught by Yonatan Bisk

Objectives of today's class

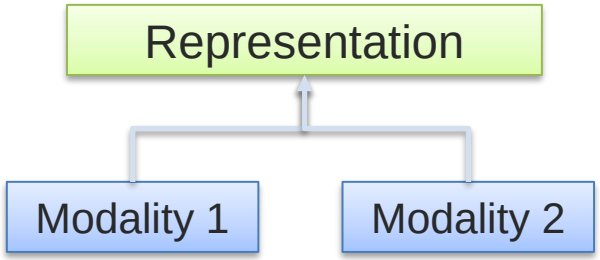
- Multi-modal representations
  - Coordinated vs. joint representations
- Unsupervised Joint representations
  - Multimodal auto-encoder
- Multi-view clustering
  - Non-negative matrix factorization
- Supervised joint representations
- Coordinated representations
  - Canonical correlation analysis
  - Deep CCA Models
  - Auto-encoder in auto-encoder

Multimodal representations

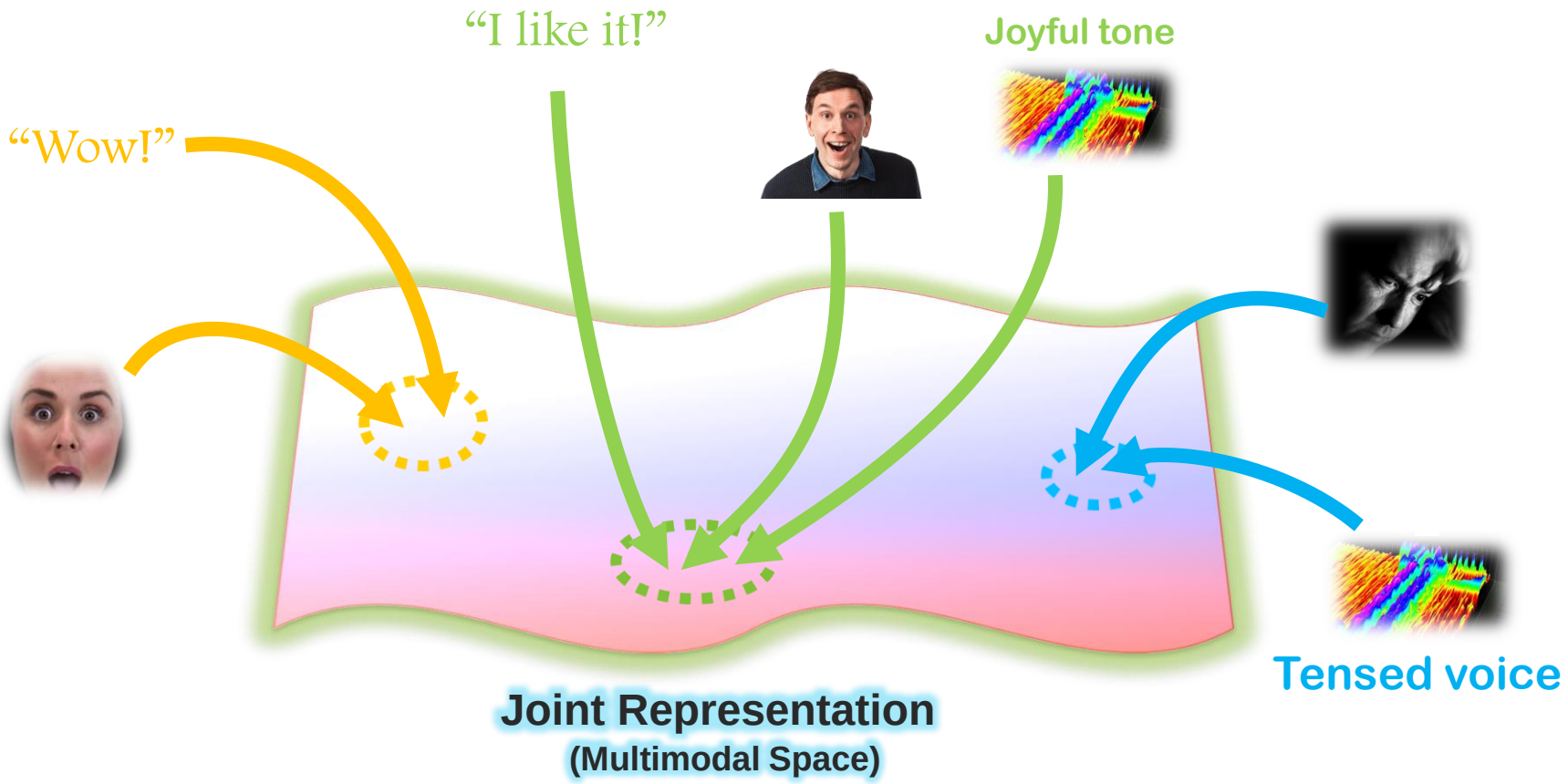
Core Challenge: Multimodal Representation

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

Ⓐ Joint representations:



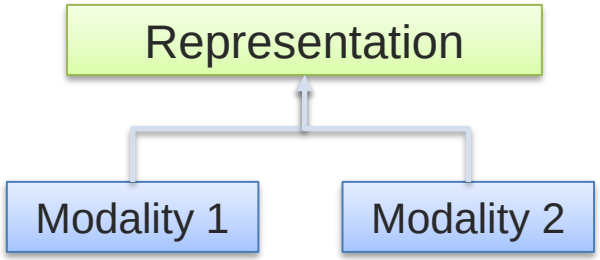
Joint Multimodal Representation



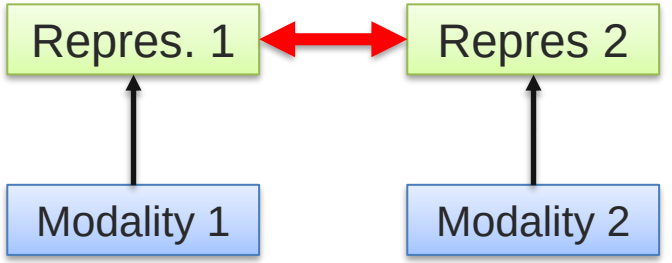
Core Challenge 1: Representation

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

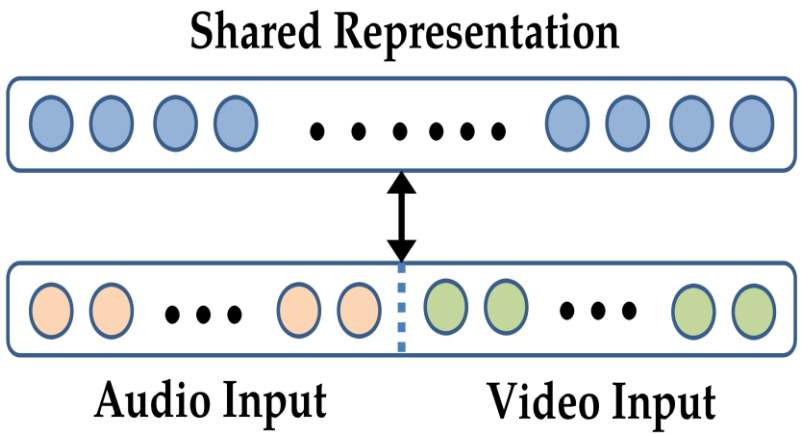
- Not just extracting features for classification
- Asking the model to be good at representing the data and not overfitting to a particular task
- Potentially allowing for better generalizability

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

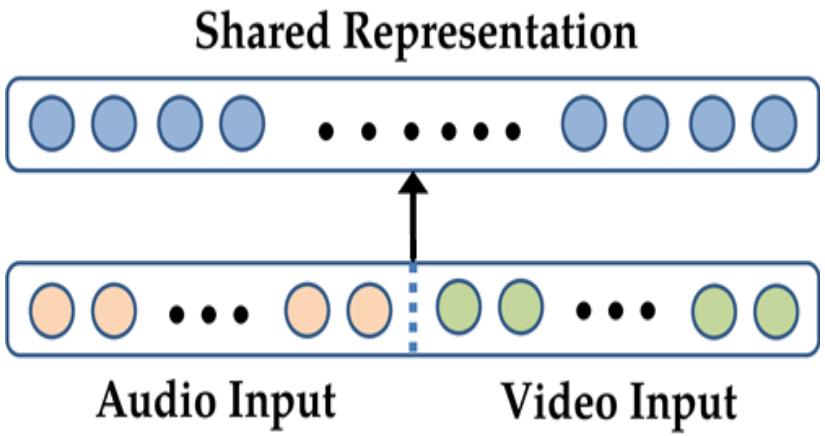
Shallow multimodal representations

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

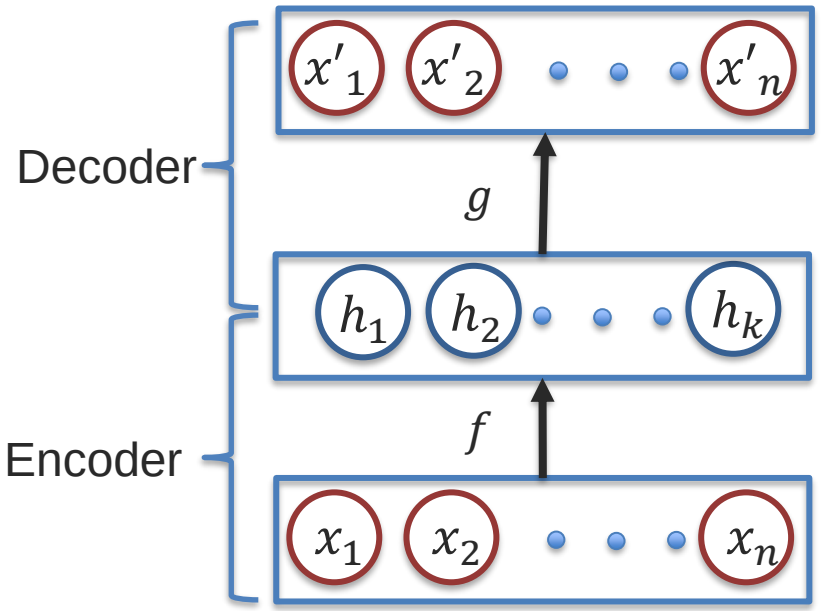
$\boldsymbol{x}' = \boldsymbol{g}(\boldsymbol{f}(\boldsymbol{x}))$ **: score function**

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

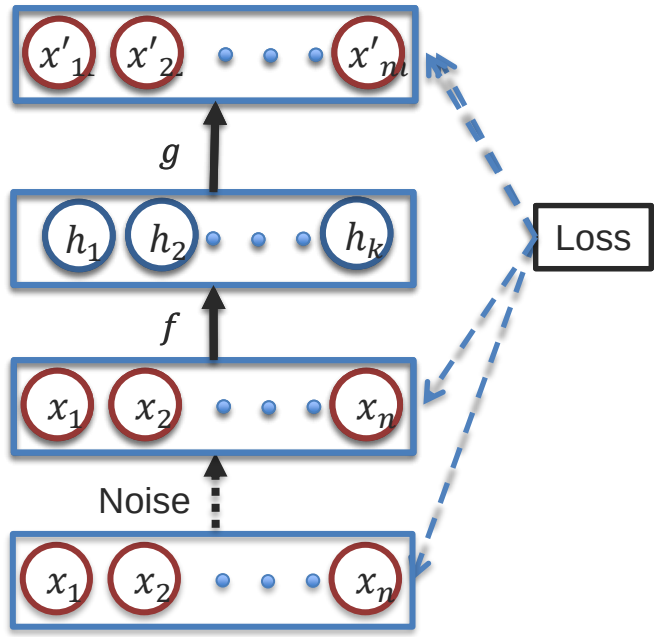
e.g., Eucledian loss: $L = \frac{1}{2}\sum_k (x_k - x'_k)^2$

But how to make it robust to noise?

Solution: Denoising autoencoder

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

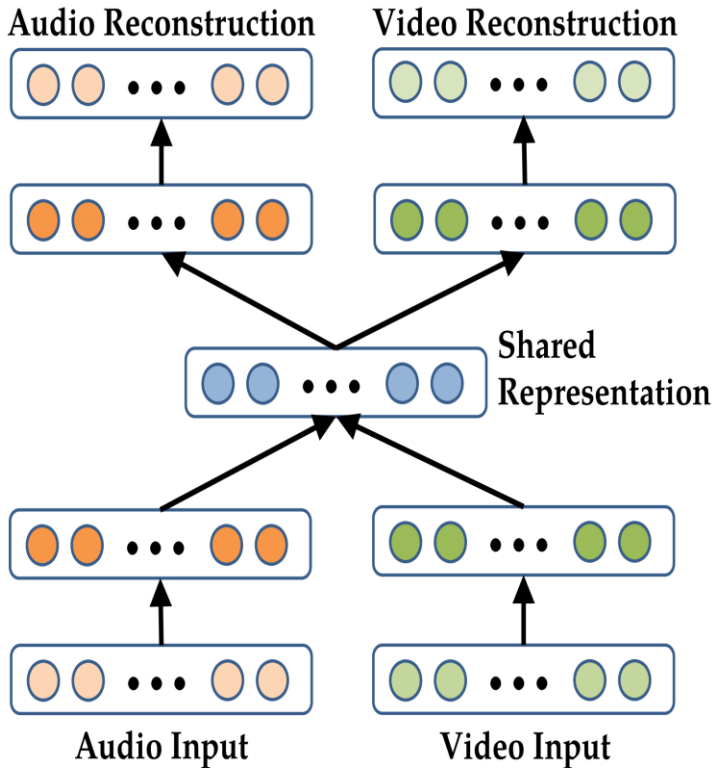
It leads to a more robust representation and prevents copying



Deep Multimodal autoencoders

Bimodal auto-encoder: a deep representation learning approach

- Used for Audio-visual speech recognition



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

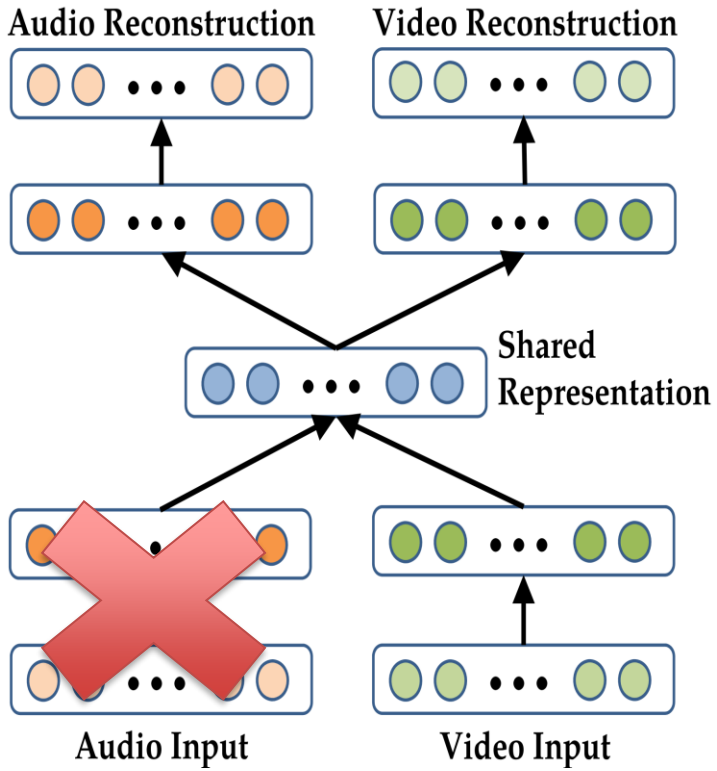
Deep Multimodal autoencoders - training

Individual modalities can be pre-trained

- Denoising Autoencoders

To train the model to reconstruct the other modality

- Use both
- Remove audio



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

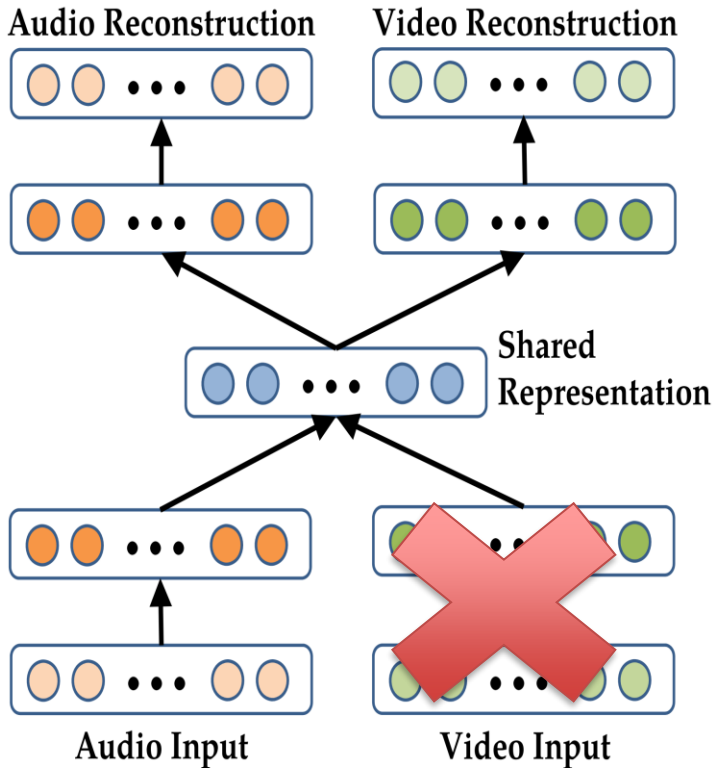
Deep Multimodal autoencoders - training

Individual modalities can be pretrained

- RBMs
- Denoising Autoencoders

To train the model to reconstruct the other modality

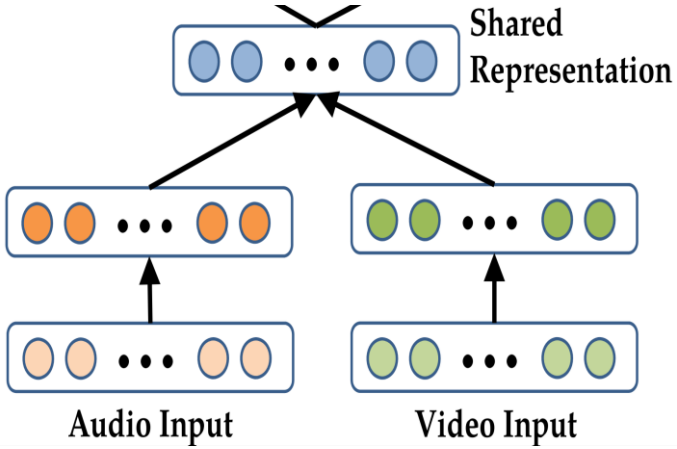
- Use both
- Remove audio
- Remove video



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

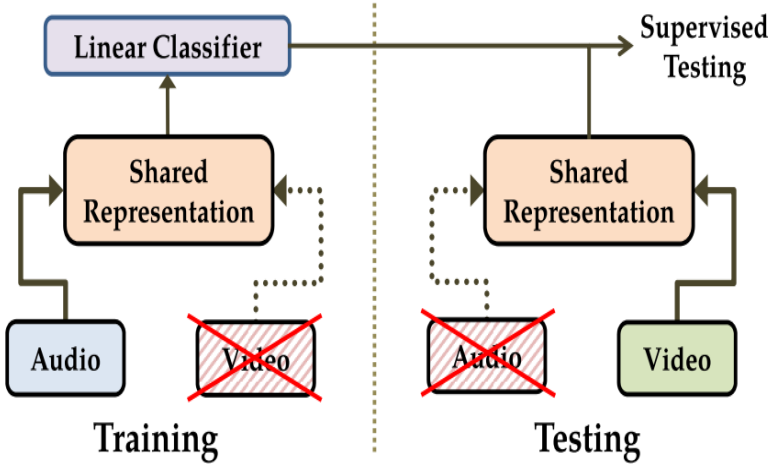
Deep Multimodal autoencoders

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



Interesting experiment:

- "Hearing to see"

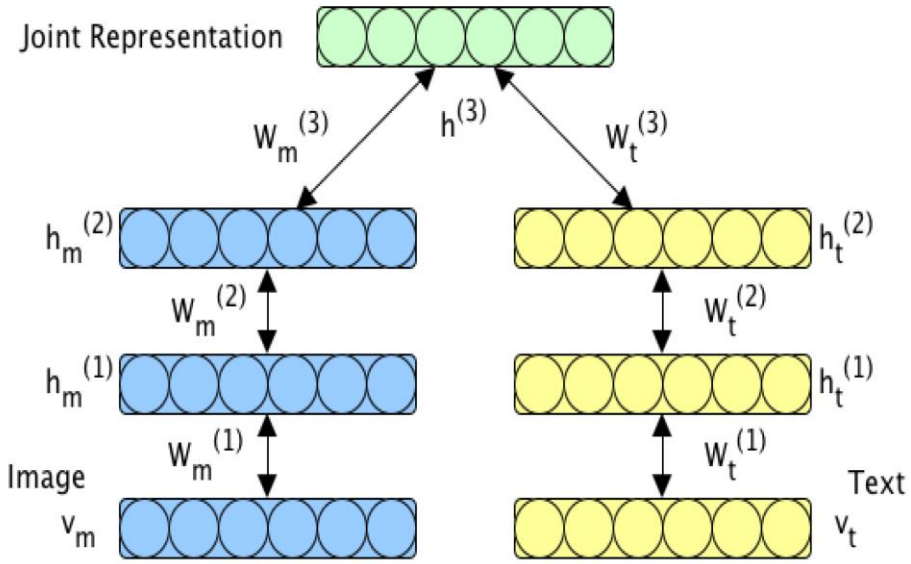


[Ngiam et al., Multimodal Deep Learning, 2011]

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]

Multi-View Clustering

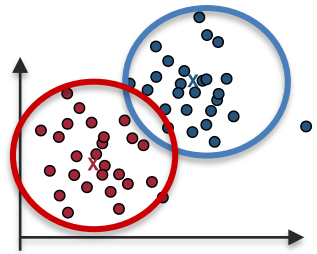
Data Clustering

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

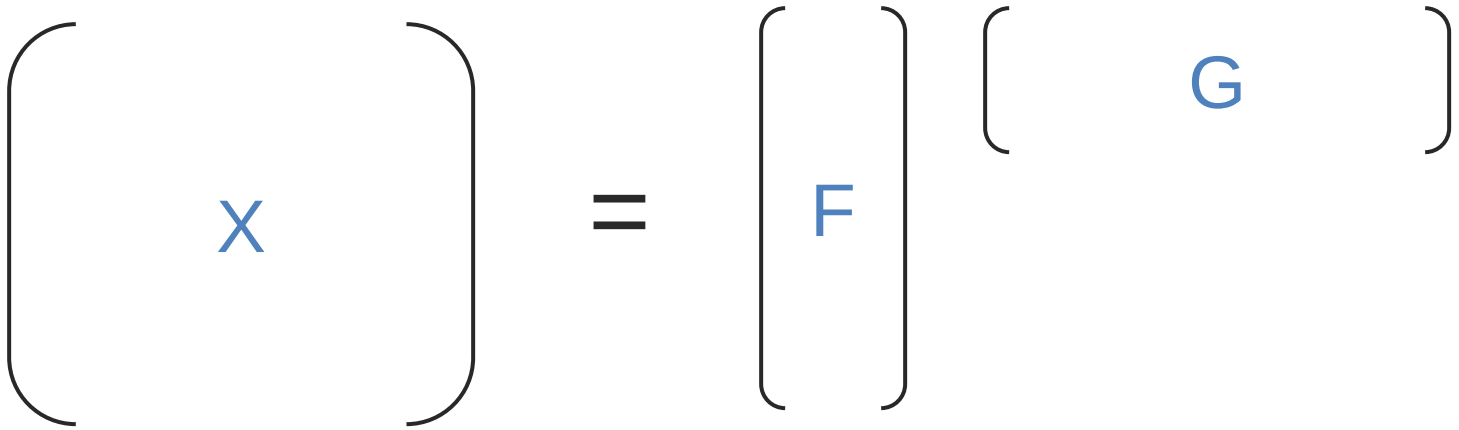
- Iterative approach
  - Assign each data point to one cluster (based on distance metric)
  - Update cluster centers
  - Until convergence
- "Winner takes all"



Image

“Soft” Clustering: Nonnegative Matrix Factorization

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



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

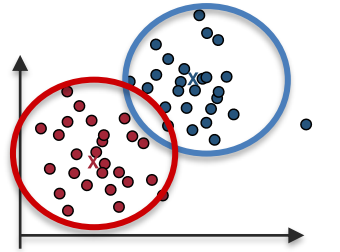
Semi-NMF and Other Extensions

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

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

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

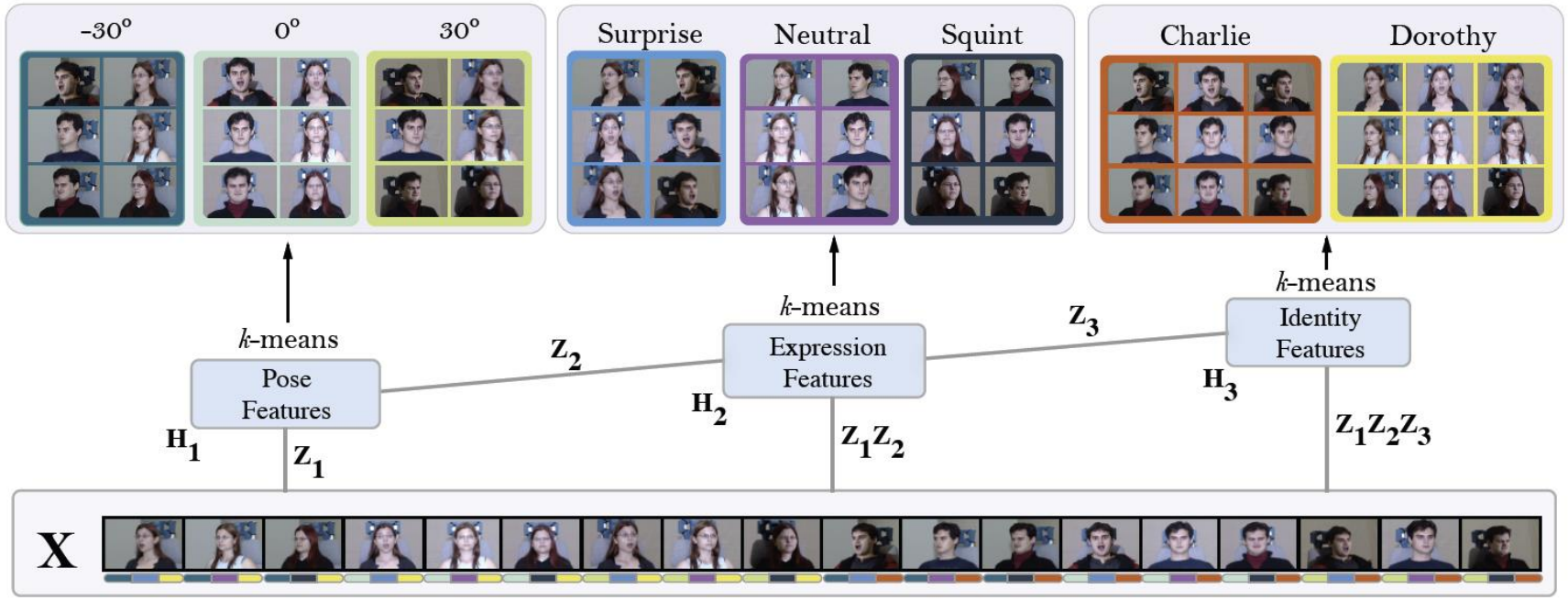
Convex-NMF: $X_{\pm} \approx X_{\pm} W_{+} G_{+}^T$



Image

Ding et al., TPAMI2015

Deep Semi-NMF Model



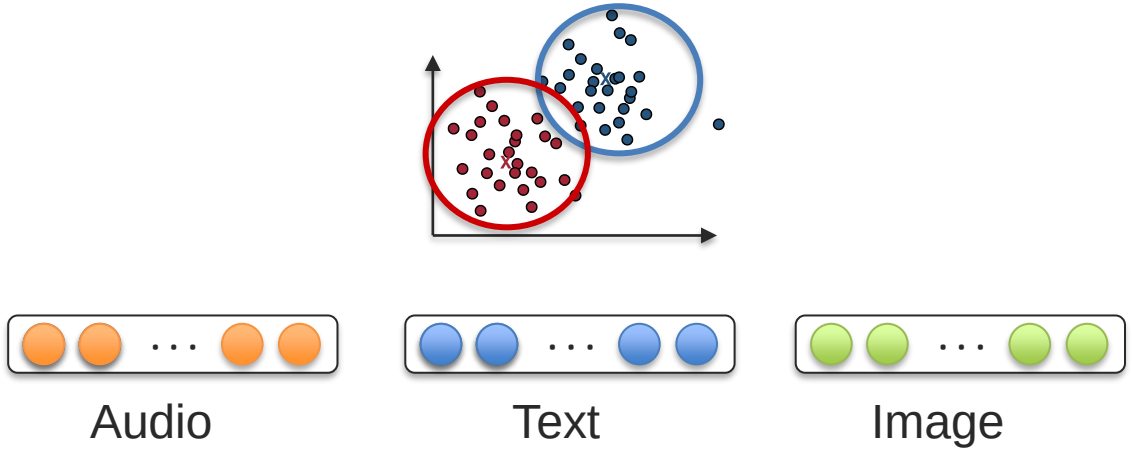
Trigerous et al., TPAMI 2015

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



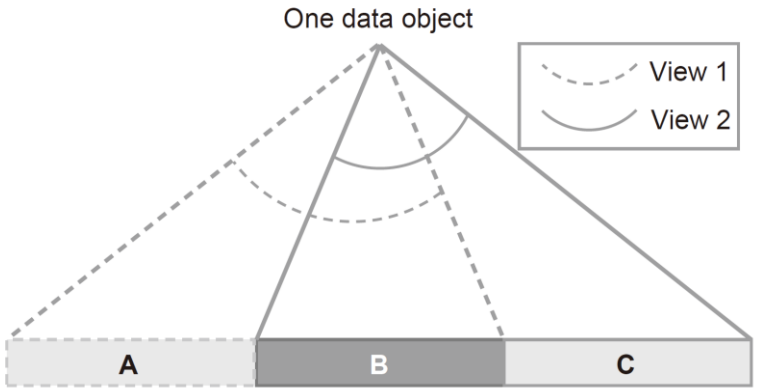
Yan Yang and Hao Wang, Multi-view Clustering: A Survey, Big data mining and analytics, Volume 1, Number 2, June 2018

Principles of Multi-View Clustering

Two important principles:

1. **Consensus principle:** maximize consistency across multiple distinct views

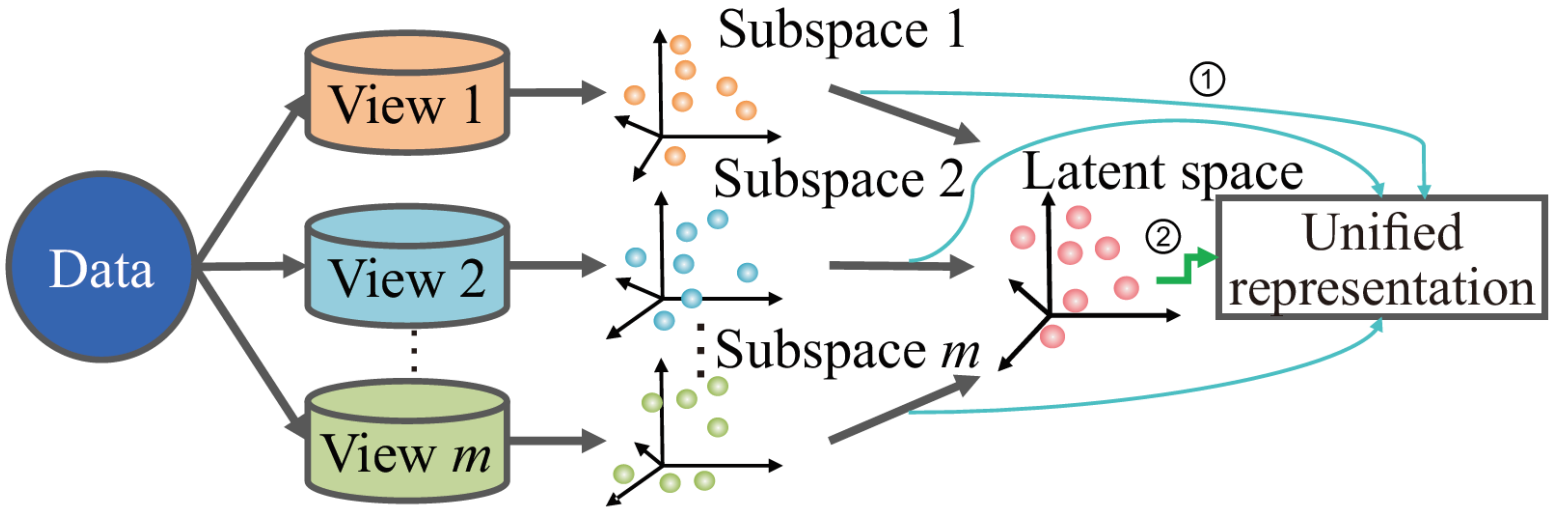
2. **Complementarity principle:** multiple views needed to get more comprehensive and accurate descriptions



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

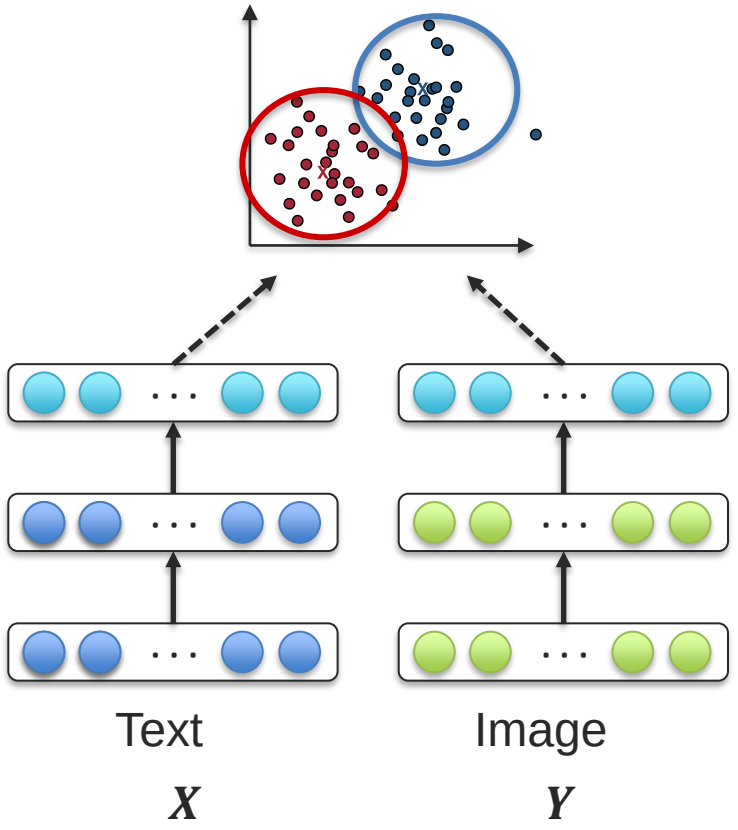
Multi-view subspace clustering

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

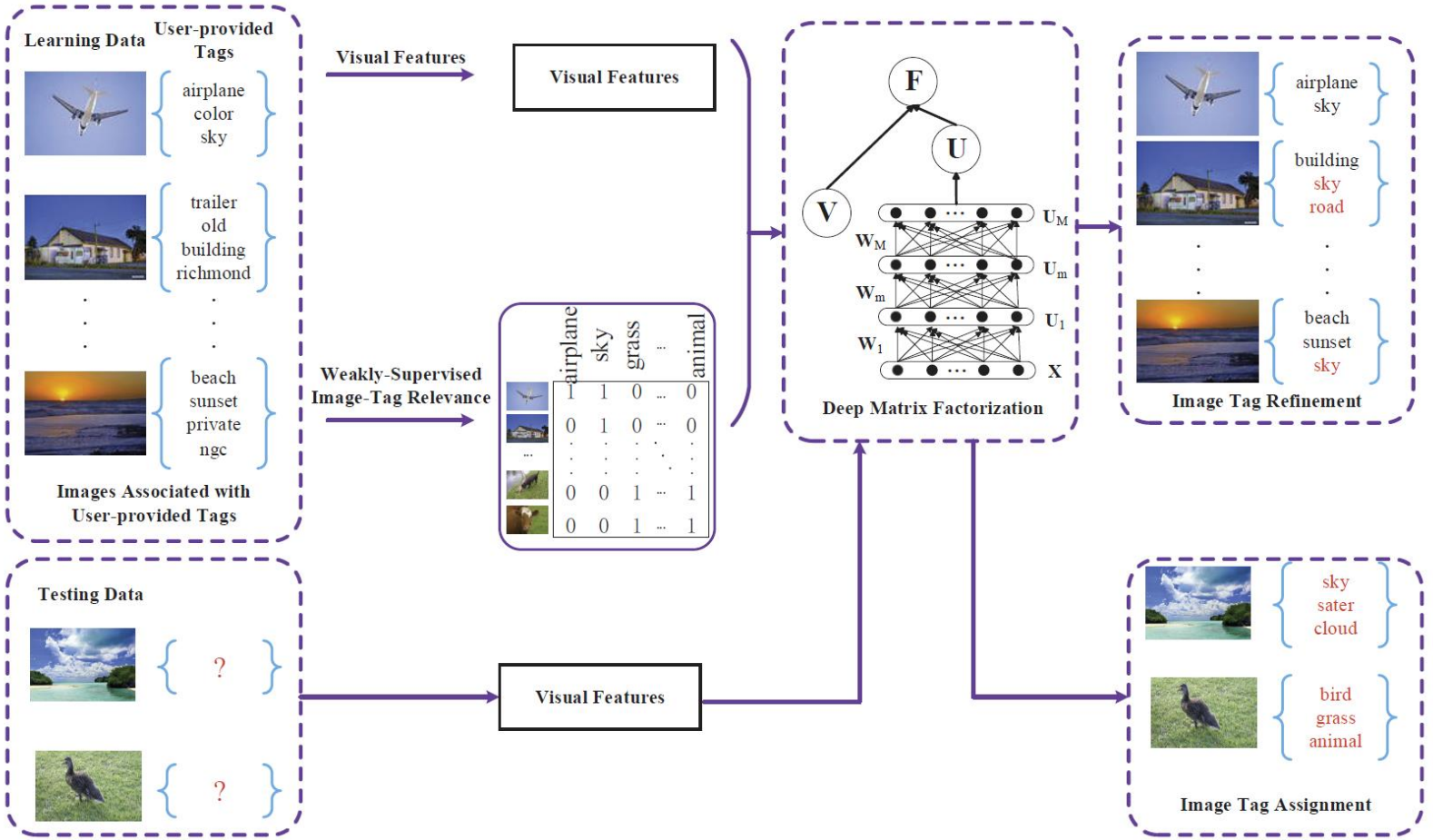


Enforcing Data Clustering in Deep Networks

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



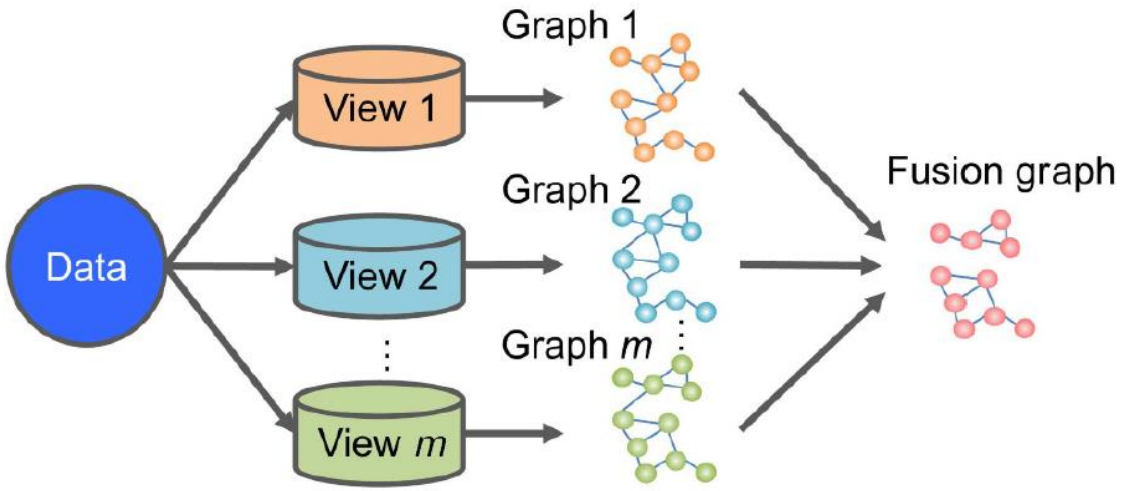
Deep Matrix Factorization



Li and Tang, MMML 2015

Other Multi-View Clustering Approaches

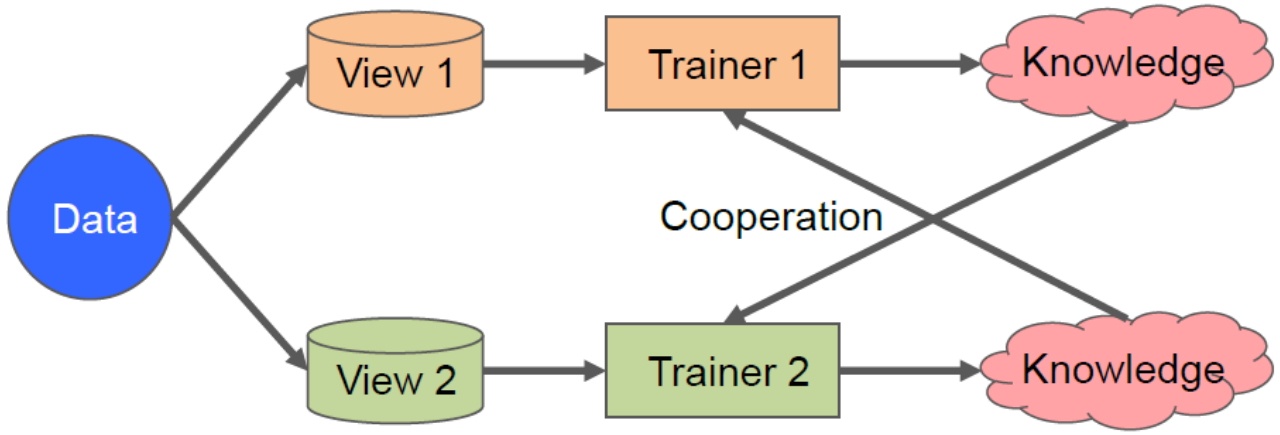
Graph-based clustering: search for a fusion graph (or network) across all views and then perform clustering



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

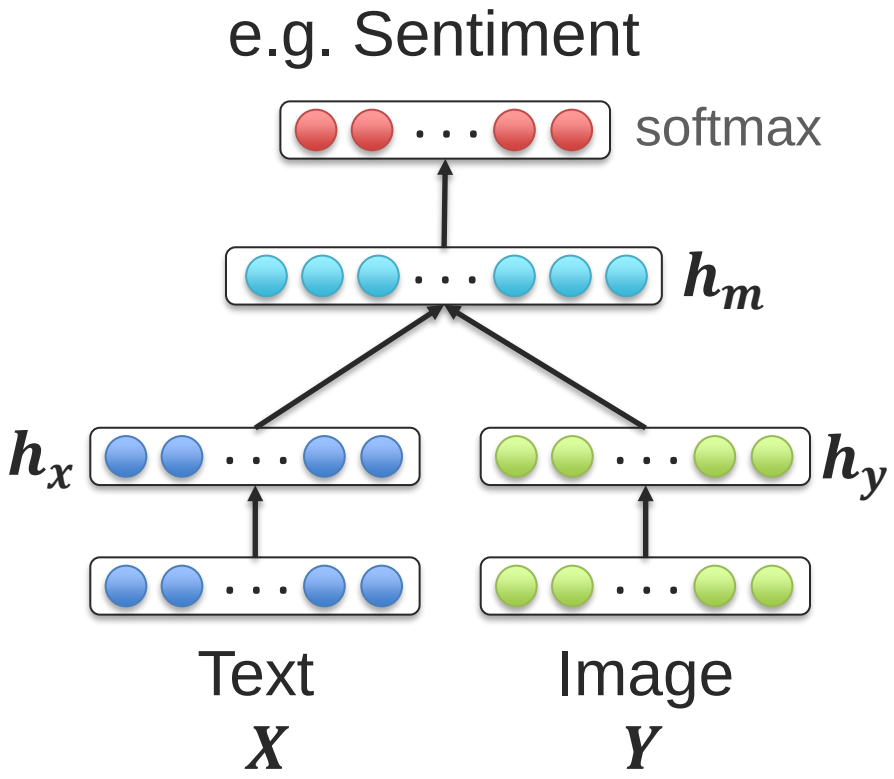
Supervised Joint representations

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?

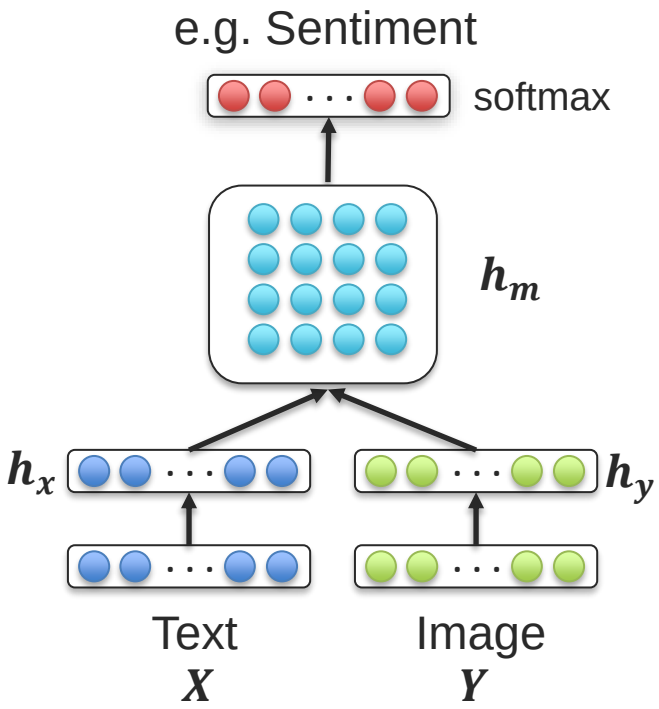


Bilinear Pooling

Models bimodal interactions:

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

[Tenenbaum and Freeman, **2000**]



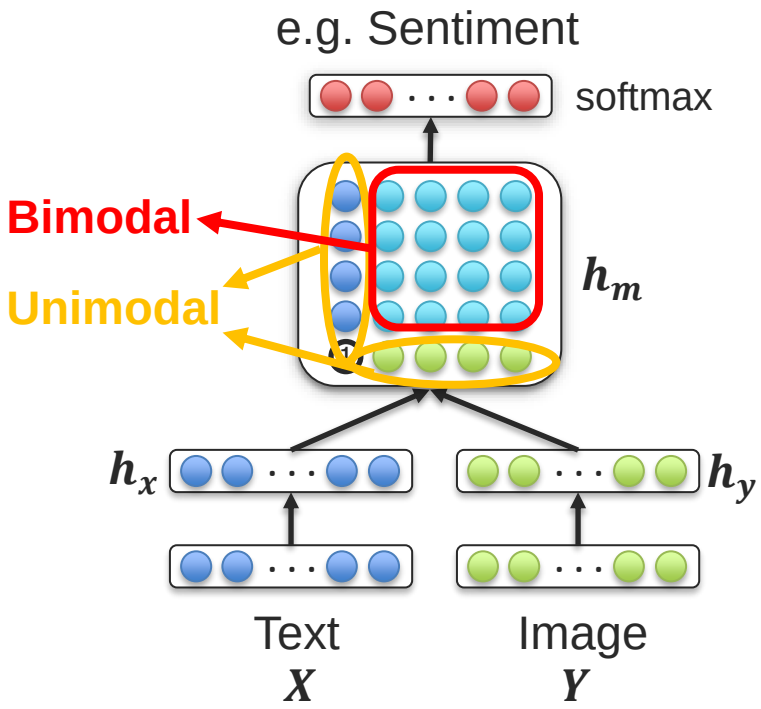
Multimodal Tensor Fusion Network (TFN)

Models both unimodal and bimodal interactions:

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

Important !

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

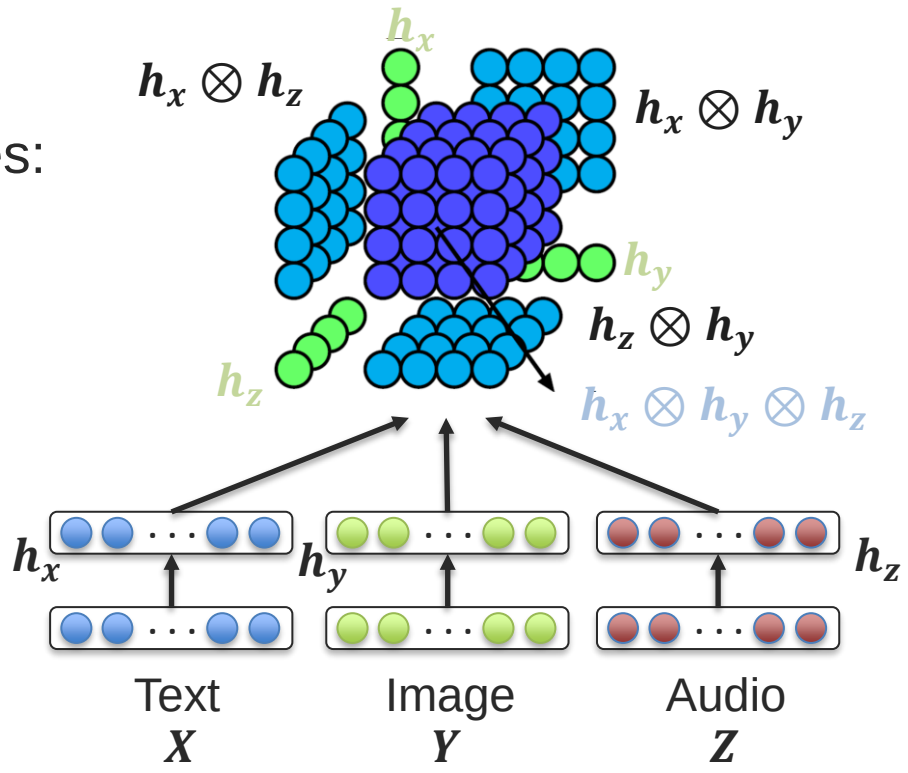


Multimodal Tensor Fusion Network (TFN)

Can be extended to three modalities:

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

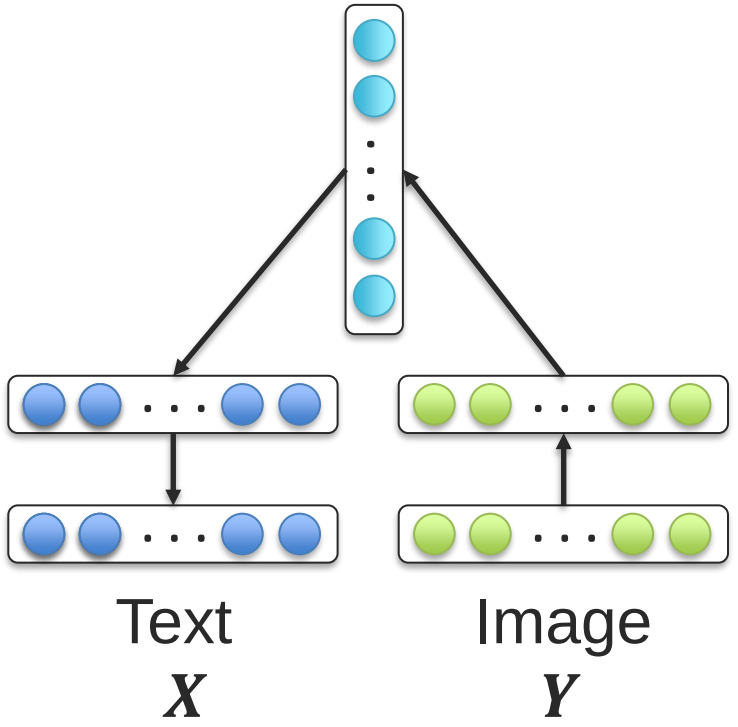
Explicitly models unimodal, bimodal and trimodal interactions !



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

Multimodal Encoder-Decoder

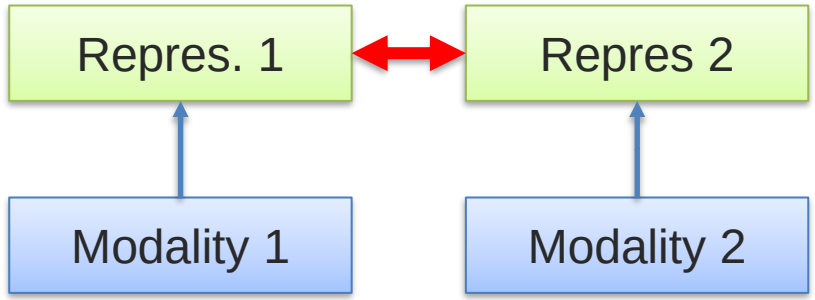
- Visual modality often encoded using CNN
- Language modality will be decoded using LSTM
  - A simple multilayer perceptron will be used to translate from visual (CNN) to language (LSTM)



Coordinated Multimodal Representations

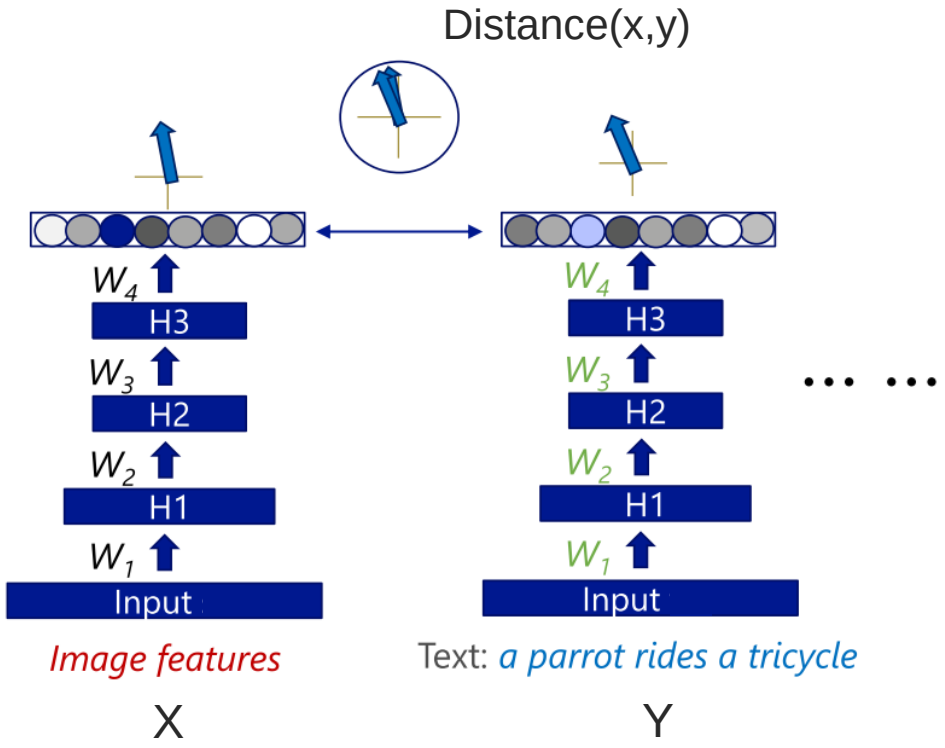
Coordinated multimodal embeddings

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



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?

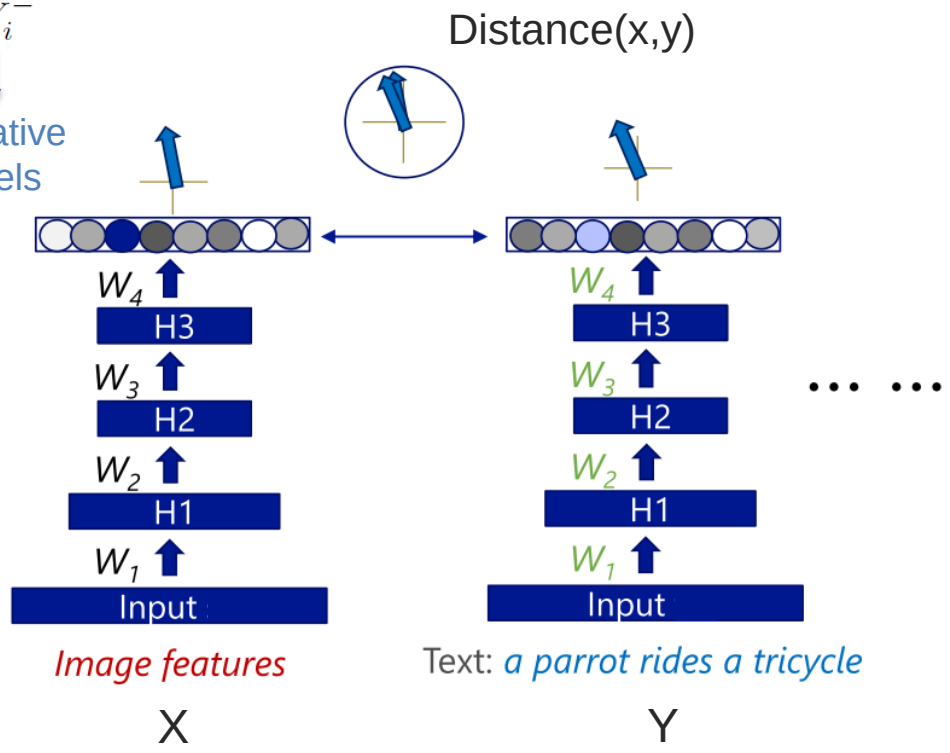
Max-margin:

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

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:

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

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

Structure-preserving constraints

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

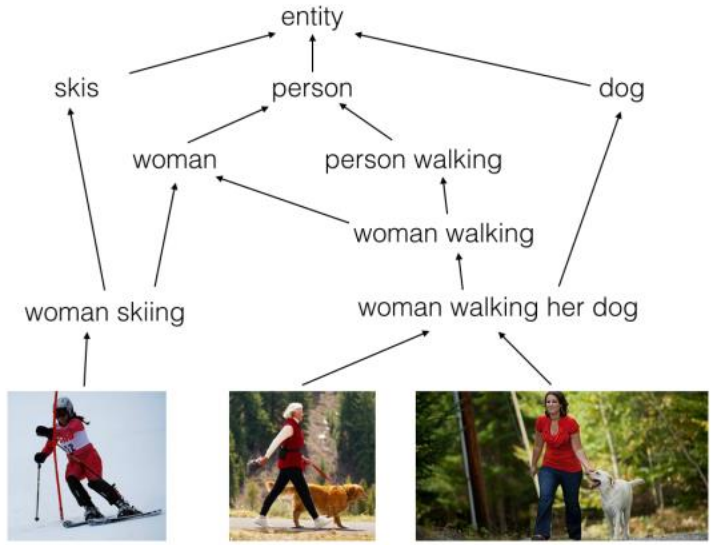
$$d(y_{i'}, y_{j'}) + m < d(y_{i'}, y_{k'}) \quad \forall y_{j'} \in N(y_{i'}), \forall y_{k'} \notin N(y_{i'})$$

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

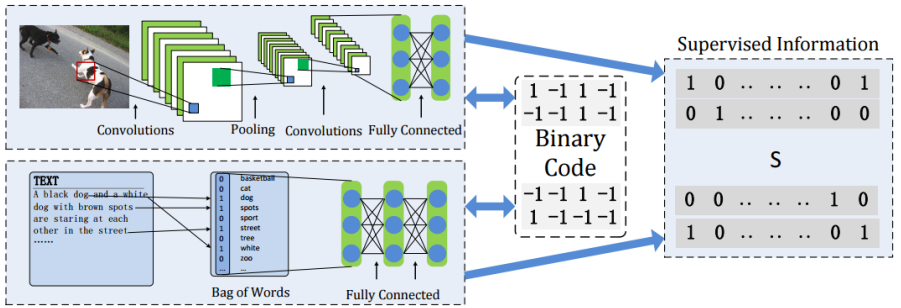
[Wang et al., Learning Deep Structure-Preserving Image-Text Embeddings, CVPR 2016]

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]

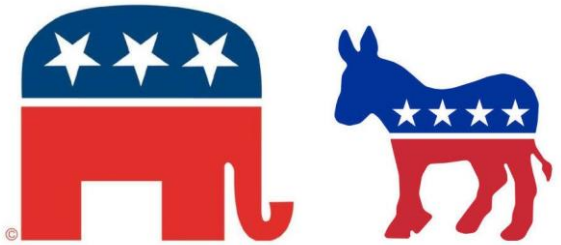
Canonical Correlation Analysis

Multi-view Learning

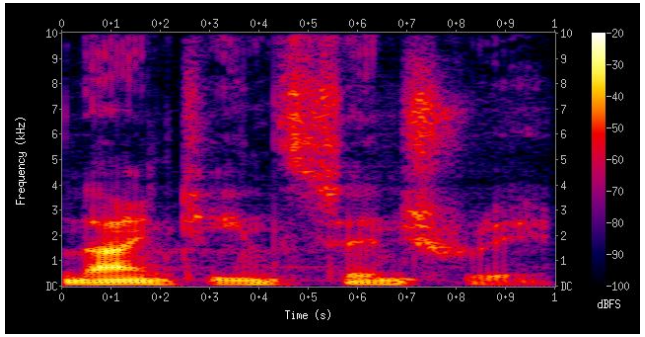
X

demographic properties

Y



responses to survey



audio features at time i

video features at time i

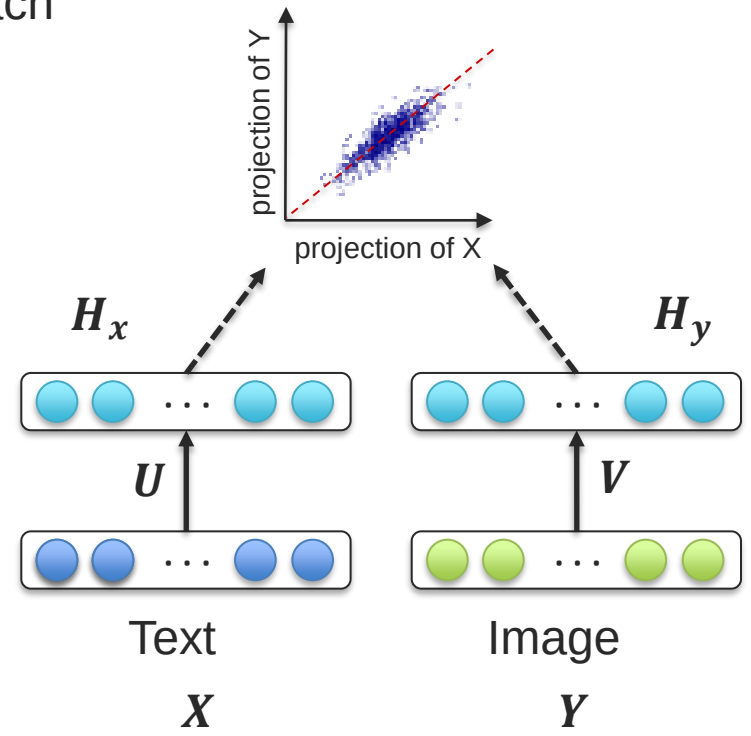
Canonical Correlation Analysis

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

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

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

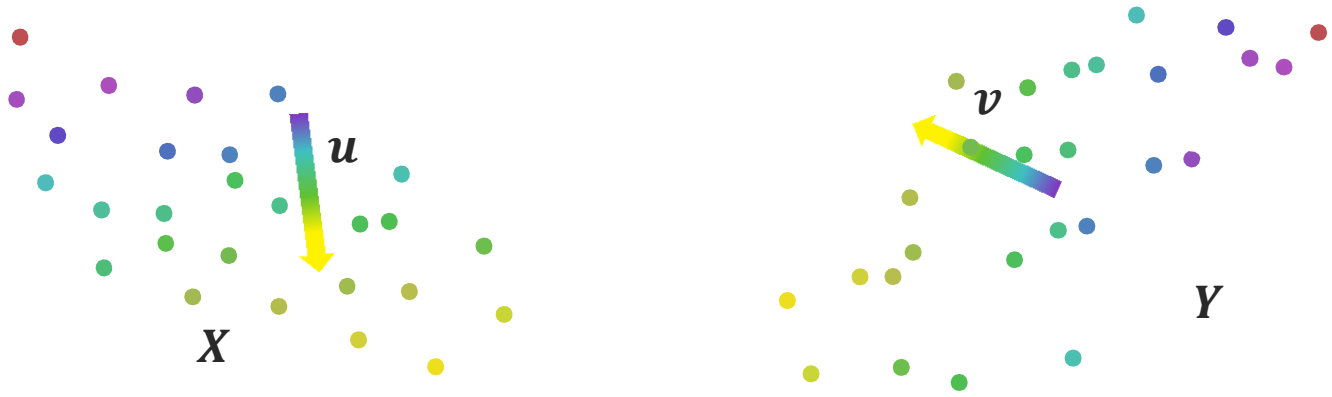
$$= \underset{\boldsymbol{u},\boldsymbol{v}}{\operatorname{argmax}}\, corr\left(\boldsymbol{u^T X}, \boldsymbol{v^T Y}\right)$$



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

Correlated Projection

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

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

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

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

Canonical Correlation Analysis

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

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

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

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

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

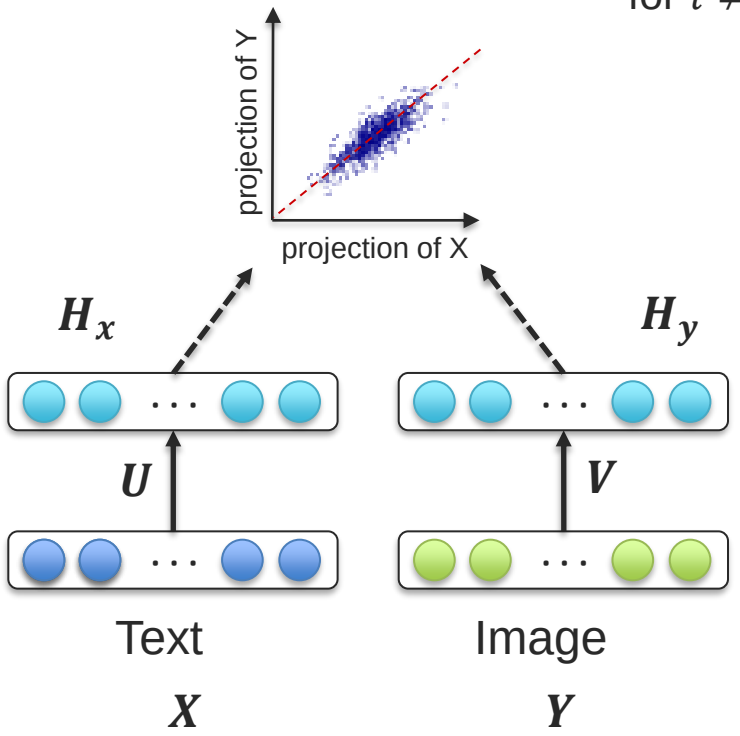
$$\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{XX}}\boldsymbol{U} = \boldsymbol{I} \qquad \boldsymbol{V}^{\boldsymbol{T}}\boldsymbol{\Sigma}_{\boldsymbol{YY}}\boldsymbol{V} = \boldsymbol{I}$$

Canonical Correlation Analysis

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

subject to: $\boldsymbol{U^T \Sigma_{XX} U = V^T \Sigma_{YY} V = I}$, $\boldsymbol{u_{(j)}^T \Sigma_{XY} v_{(i)} = 0}$ 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

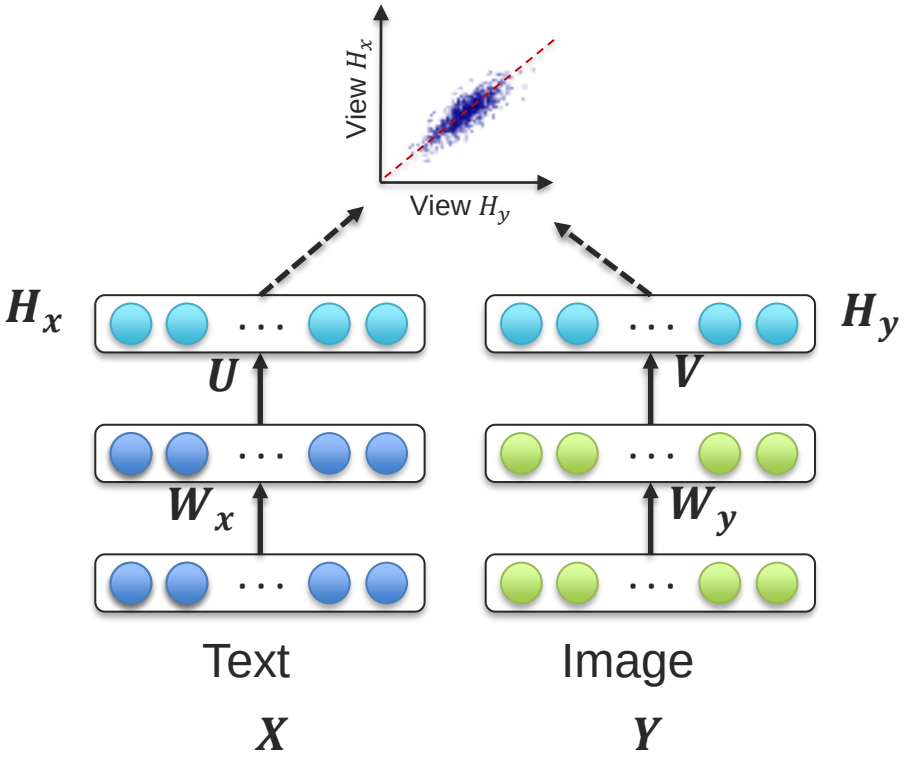
Same objective function as CCA:

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

And need to compute gradients:

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

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

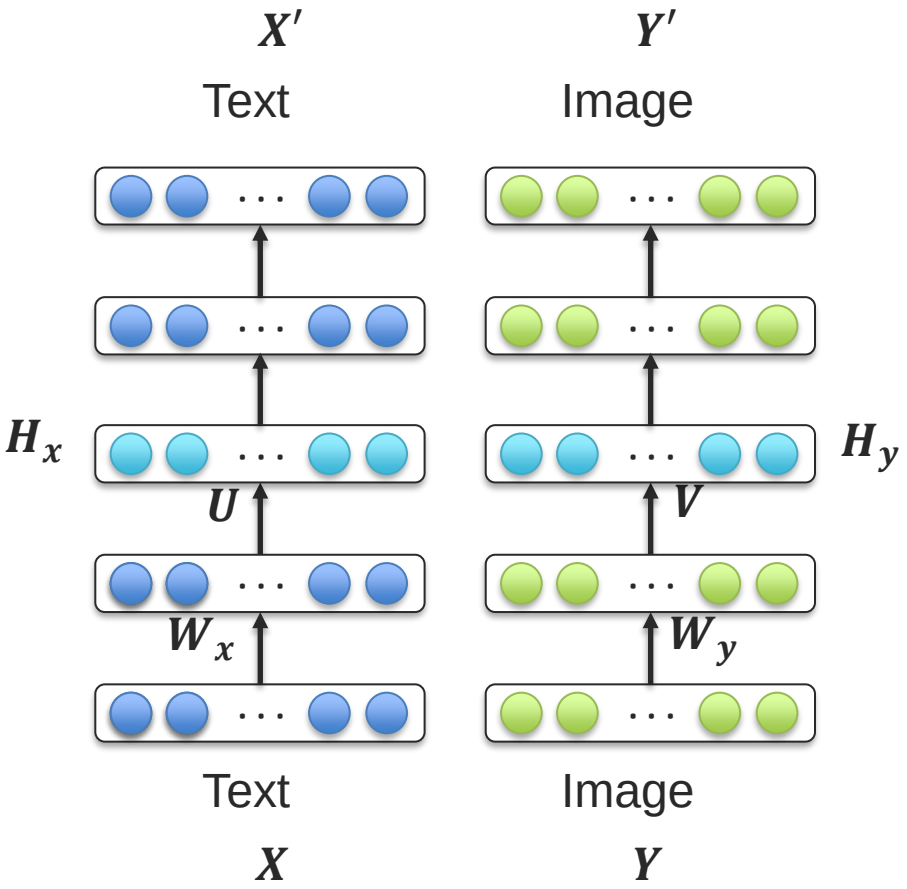


Andrew et al., ICML 2013

Deep Canonical Correlation Analysis

Training procedure:

1. Pre-train the models parameters using denoising autoencoders

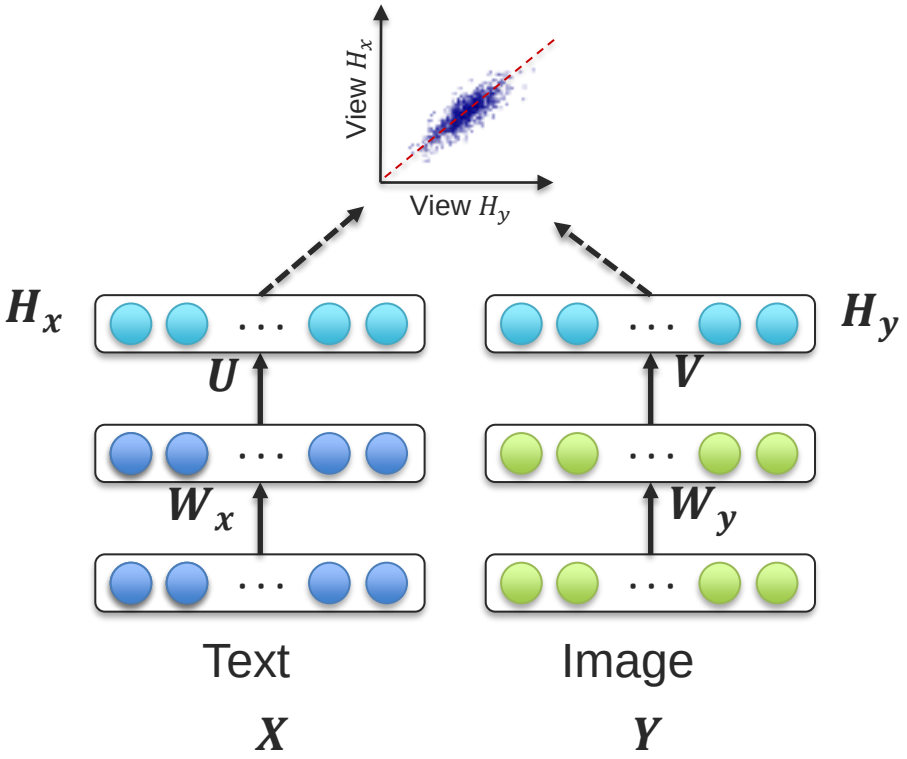


Andrew et al., ICML 2013

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)

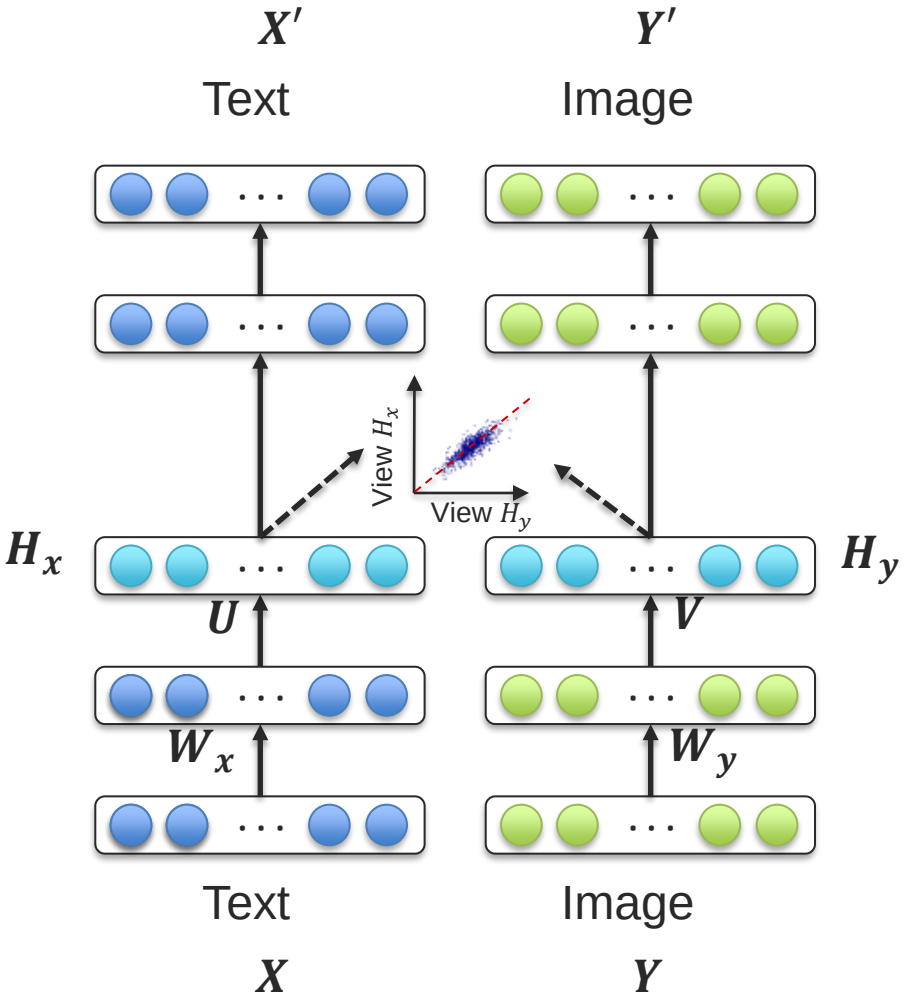


Andrew et al., ICML 2013

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

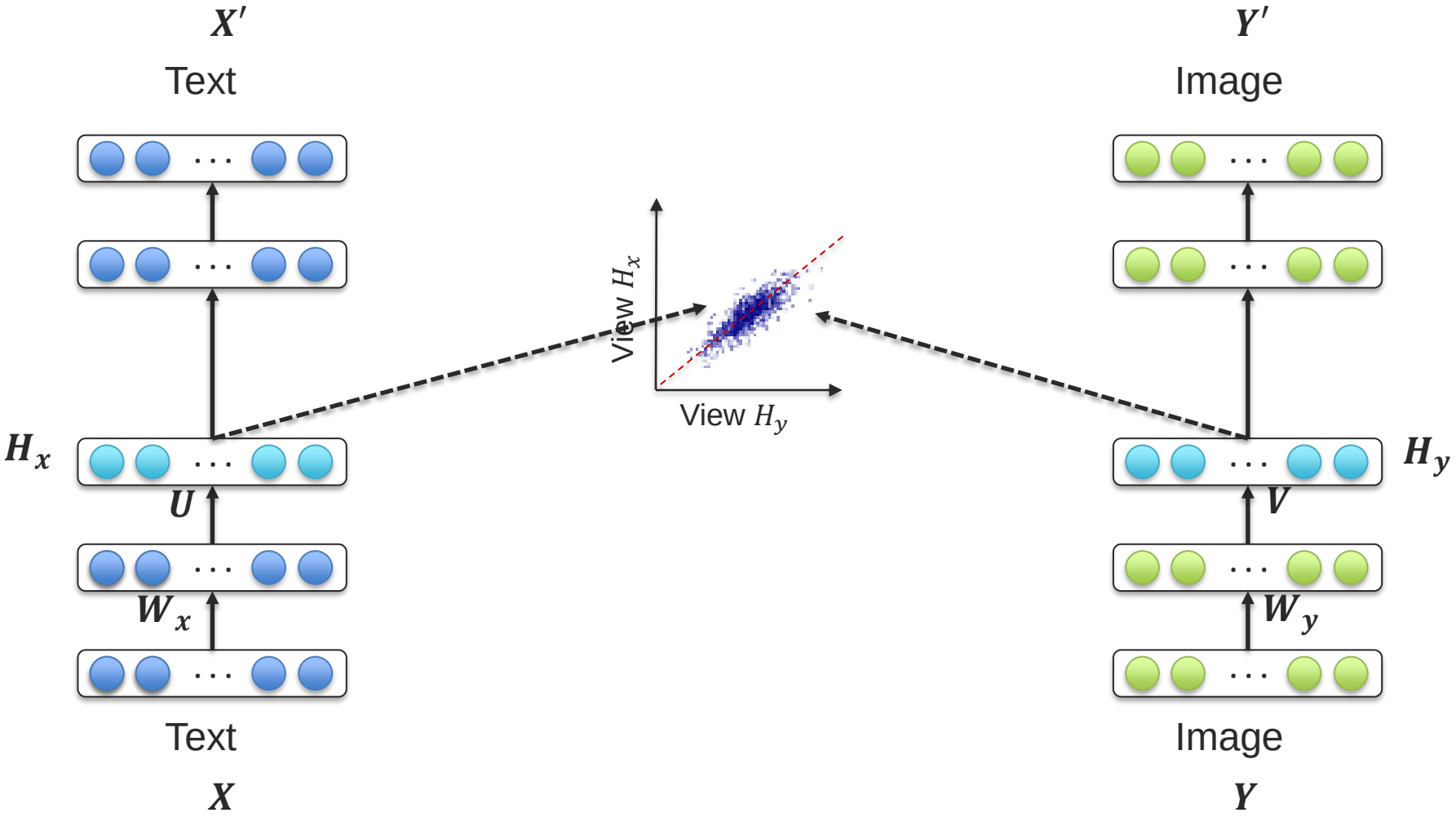


Wang et al., ICML 2015

Auto-Encoder in Auto-Encoder Network

Deep Canonically Correlated Autoencoders (DCCAE)

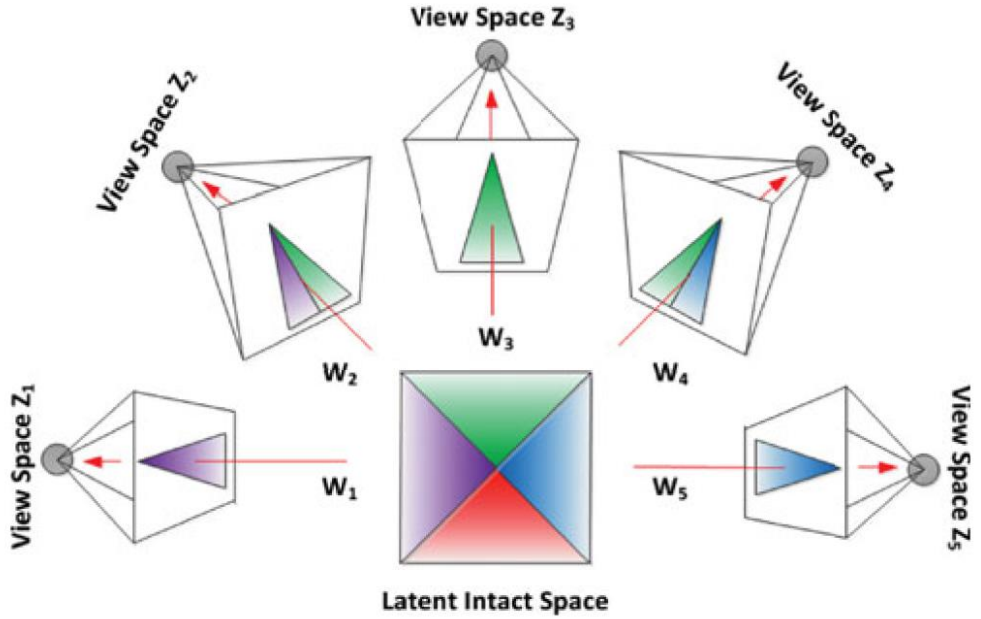
Wang et al., ICML 2015



Multi-view Latent "Intact" Space

Xu et al., TPAMI 2015

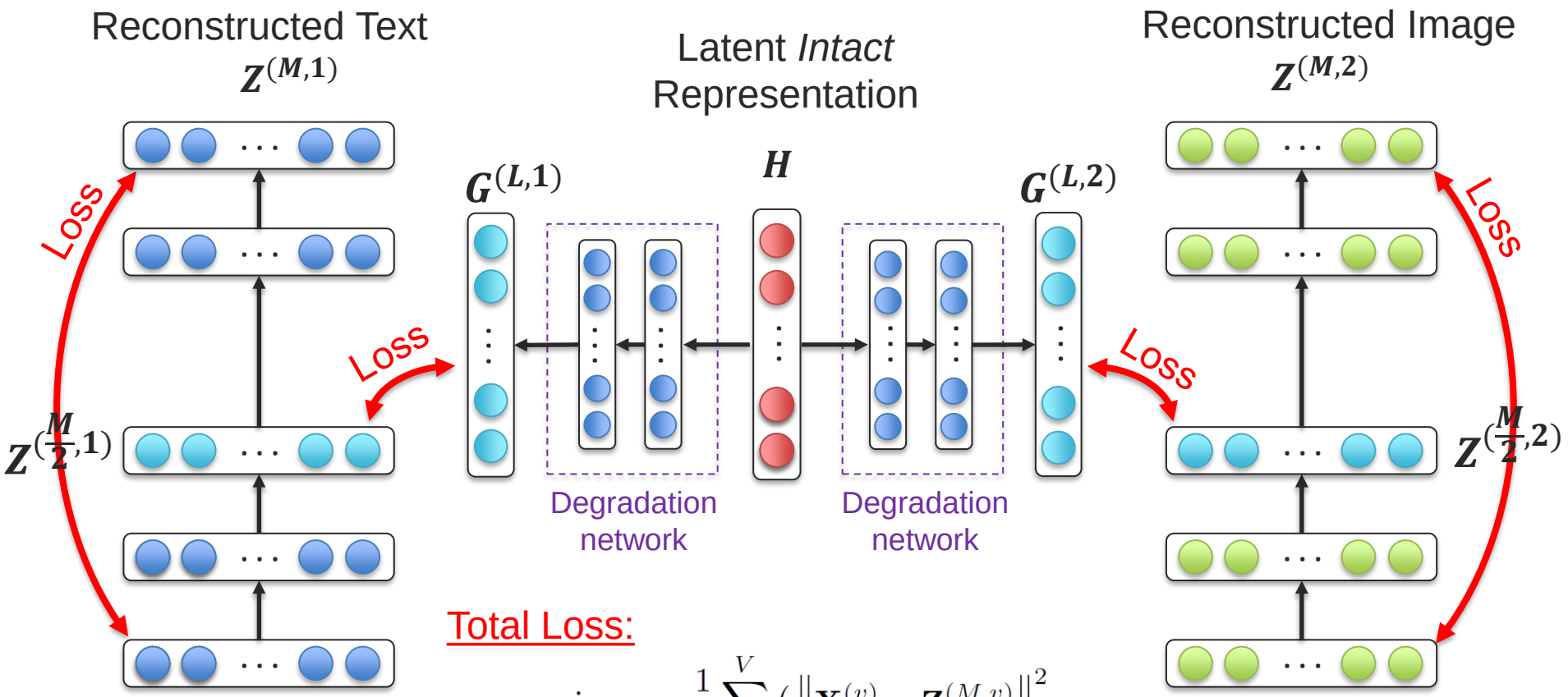
Given multiple views z_i from the same "object":



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