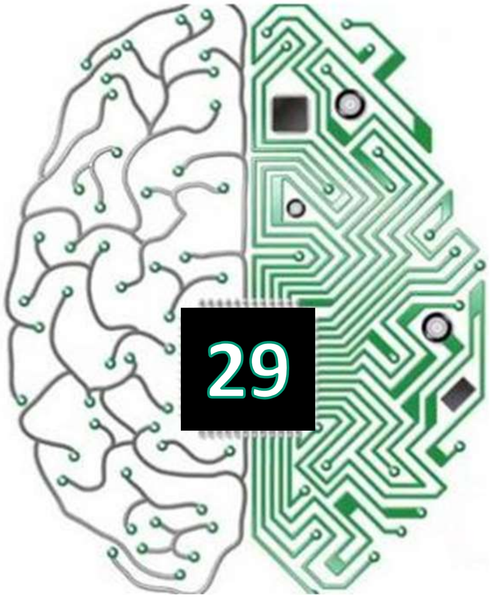
Open Elective Course [OE]

Course Code: CSO507

Winter 2023-24

Lecture#



Deep Learning

Unit-6: Representation Learning (Part II)

Course Instructor:

Dr. Monidipa Das

Assistant Professor

Department of Computer Science and Engineering

Indian Institute of Technology (Indian School of Mines) Dhanbad, Jharkhand 826004, India

PCA: Algorithmic View in Detail

- Consider set of data points $\{x_i\}$ where $i = 1, \dots, N$ and $x_i \in \mathbb{R}^D$
  – Mean of the original data: $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

- Goal: Project data onto $K < D$ dimensional space while maximizing the variance of the projected data

- To begin with, consider $K = 1$:
  – Let w_1 be the direction of the projection.
  – Set $||w_1|| = 1$, as it is only the direction that is important
  – Projected data: $w_1^T x_i$ and Projected mean: $w_1^T \bar{x}$

PCA: Algorithmic View in Detail

- Covariance of original data:

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

- Variance of the projected data:

$$\frac{1}{N}\sum_{i=1}^{N}\left\{w_1^T x_i - w_1^T \bar{x}\right\}^2 = \frac{1}{N}\sum_{i=1}^{N}\left\{w_1^T (x_i - \bar{x})\right\}^2$$

$$\frac{1}{N}\sum_{i=1}^{N}\left\{w_1^T (x_i - \bar{x})(x_i - \bar{x})^T w_1\right\} \quad = w_1^T \Sigma w_1$$

- Goal: maximizing variance of the projected data:

$$\max_{w_1} \; w_1^T \Sigma w_1 \text{ such that } ||w_1|| = 1$$

PCA: Algorithmic View in Detail

- Using Lagrange multipliers

$$\max_{w_1} \quad w_1^T \Sigma w_1 + \lambda_1 (1 - w_1^T w_1)$$

- By setting the derivative w. r. t. w_1 equal to 0

$$\Sigma w_1 = \lambda_1 w_1$$

  - w_1 must be an ***eigenvector*** of Σ
  - the variance is maximized by choosing the eigenvector associated with the largest eigenvalue.

- w_1 corresponds to the first principal component.

PCA: Algorithm

1. Create $N \times D$ data matrix X, with one row vector x_i per data point
2. **Subtract mean $\bar{x}$ from each row vector x_i in X**
3. $\Sigma \leftarrow$ covariance matrix of X
4. Find eigenvectors W and eigenvalues Λ of Σ
5. Principal Components W_K are the K eigenvectors with largest eigenvalues
6. Transformed data $\underbrace{Y}_{N \times K} = \underbrace{X}_{N \times D}\underbrace{W_K}_{D \times K}$

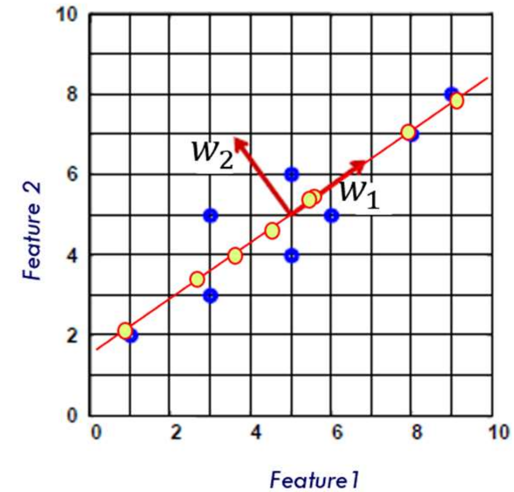
$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

PCA: Example

- Compute the principal components for the following two-dimensional dataset

$$X = \{(1,2), (3,3), (3,5), (5,4), (5,6), (6,5), (8,7), (9,8)\}$$



- SOLUTION
  - Mean-centering the data: $\bar{x} = (5,5)$
    $\{(-4,-3), (-2,-2), (-2,0), (0,-1), (0,1), (1,0), (3,2), (4,3)\}$
  - The covariance estimate of the data is: $\Sigma_x = \begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}$
  - Estimation of eigenvalues $\Sigma_x w = \lambda w \Rightarrow |\Sigma_x - \lambda I| = 0 \Rightarrow \begin{vmatrix} 6.25-\lambda & 4.25 \\ 4.25 & 3.5-\lambda \end{vmatrix} = 0 \Rightarrow \lambda_1 = 9.34;\ \lambda_2 = 0.41;$
  - The eigenvectors are the solutions of the system

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}\begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} = \begin{bmatrix} \lambda_1 w_{11} \\ \lambda_1 w_{12} \end{bmatrix} \Rightarrow \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$$

$$\begin{bmatrix} 6.25 & 4.25 \\ 4.25 & 3.5 \end{bmatrix}\begin{bmatrix} w_{21} \\ w_{22} \end{bmatrix} = \begin{bmatrix} \lambda_2 w_{21} \\ \lambda_2 w_{22} \end{bmatrix} \Rightarrow \begin{bmatrix} w_{21} \\ w_{22} \end{bmatrix} = \begin{bmatrix} -0.59 \\ 0.81 \end{bmatrix}$$

  - Transformed data $W_K = \begin{bmatrix} w_{11} \\ w_{12} \end{bmatrix} = \begin{bmatrix} 0.81 \\ 0.59 \end{bmatrix}$

$$Y = XW_K = \{(-5.0), (-2.8), (-1.6), (-0.6), (0.6), (0.8), (3.6), (5.0)\}$$

PCA: How to choose K?

- Choose K using the following criterion:

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{D} \lambda_i} > Threshold\ (e.g.\ 0.90\ or\ 0.95)$$

- In this case, we say that we “preserve” 90% or 95% of the information (variance) in the data.

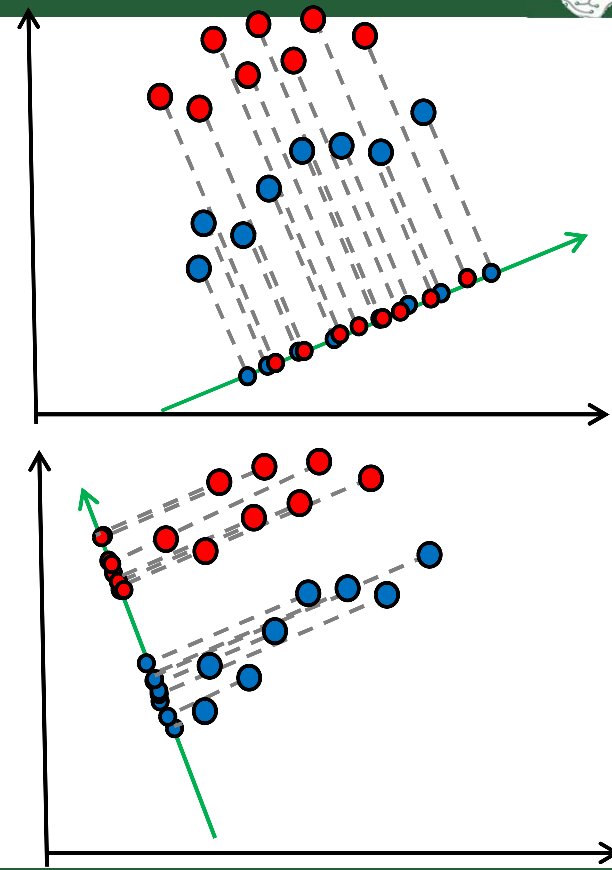
- If $K = D$, then we “preserve” 100% of the information in the data.

PCA: Benefits

- PCA identifies the strongest patterns in the data in an unsupervised way
- Capture most of the variability of the data by a small fraction of the total set of dimensions
- Eliminate much of the noise in the data, making it beneficial for various learning algorithms

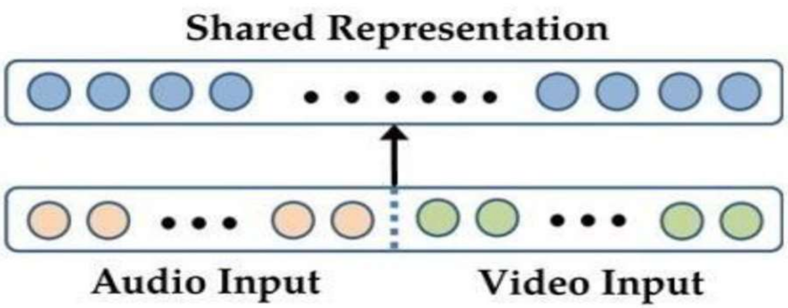
PCA: Limitations

- What if very large dimensional data?
  - D=10^4 → $|\Sigma| = 10^8$
- PCA does not consider class separability since it does not take into account the class label of the feature vector
- PCA simply performs a coordinate rotation that aligns the transformed axes with the directions of maximum variance
- There is no guarantee that the directions of maximum variance will contain good features for discrimination



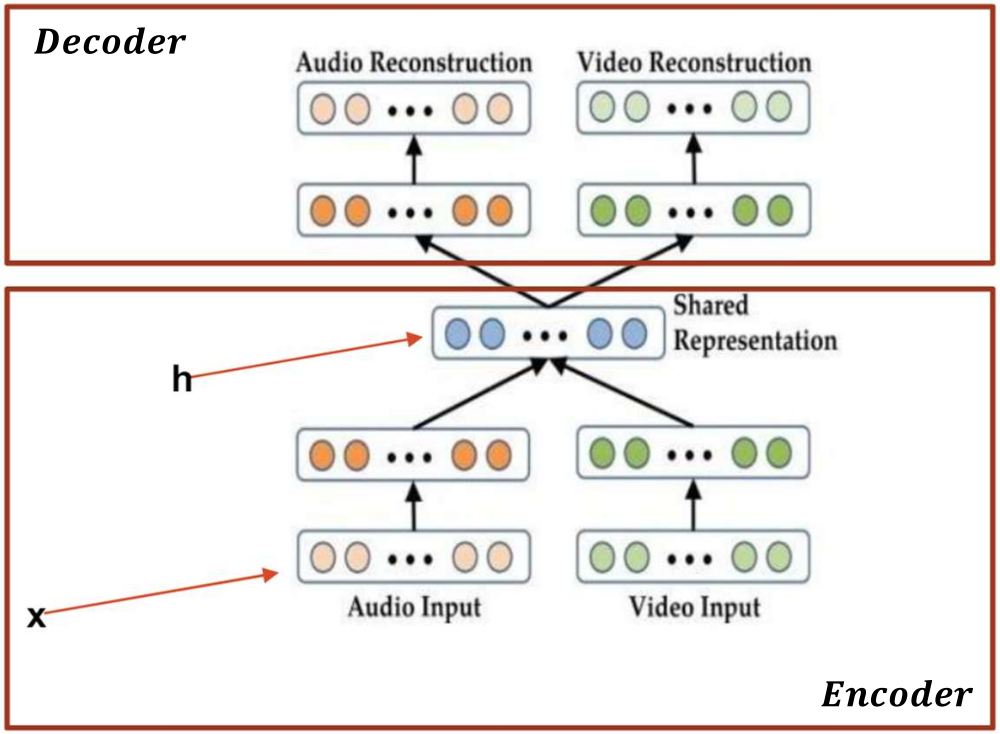
Multimodal Representation Learning

- How do we deal with tasks involving 2 or more modalities?
- For instance, given an image and a question about it, find the answer OR VQA.
  - Approach: Simply concatenate representations and plug that in your end-to-end network.



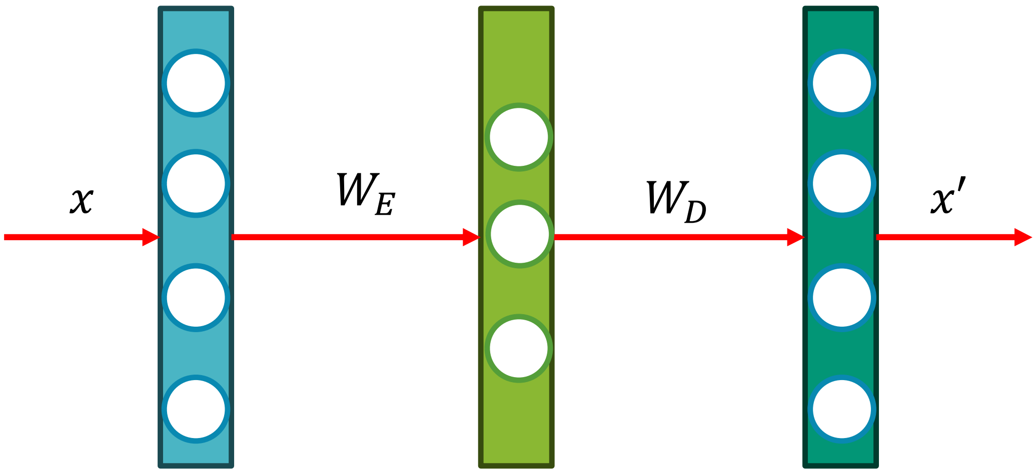
Autoencoders

- Encode modalities in a shared space
- Train and then when training the downstream task keep only the encoder part
  - Pros : Extremely robust, can reconstruct missing modalities if trained well
  - Cons : Needs separate training, and often not state-of-the-art compared to pooled or coordinated representations

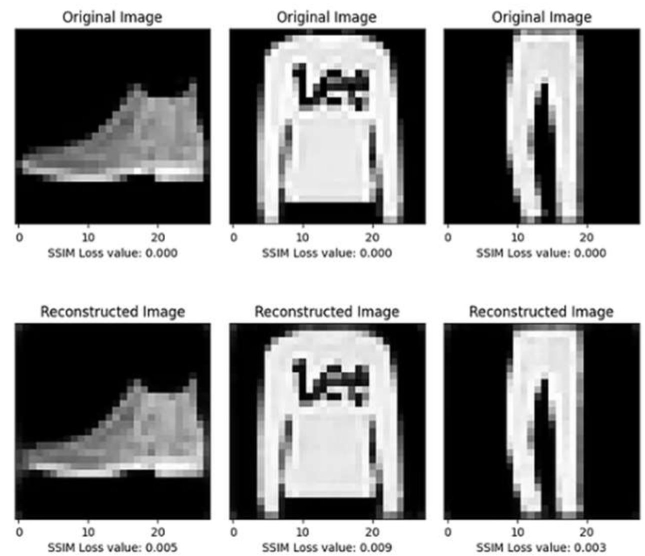


Autoencoders

- Basic Architecture



$$h = g(x.W_E + b_E) \qquad x' = g(h.W_D + b_D)$$



Autoencoders

- Choice of Loss Function:
  - Case-1: Binary Input

$$\mathcal{L} = \sum_{j=1}^{m} \boxed{\sum_{i=1}^{n} C(p_{ij}, q_{ij})}$$

$$-\sum_{i=1}^{n} p_i * \log(q_i) + (1 - p_i) * \log(1 - q_i)$$

  - Case-2: Real Input

$$\mathcal{L} = \sum_{j=1}^{m} \sum_{i=1}^{n} (x_{ij} - \hat{x}_{ij})^2$$

Autoencoder and PCA

- Autoencoders can be used to perform like PCA
  - Standardized input
  - Hidden Layer: Linear activation
  - Output Layer: Linear activation
  - Loss Function: Mean-squared Error

- Variability structure may not always be linear

Types of Autoencoders

- Sparse Autoencoder
- Contractive Autoencoder
- Denoising Autoencoder
- Variational Autoencoder
- ……

To be discussed in the next unit….

Questions?