Multimodal Deep Learning With An Application To Folk Art Recommendation.

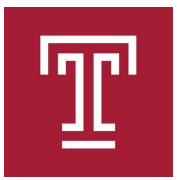
Haimonti Dutta

Department of Management Science and Systems
The State University of New York, Buffalo
New York, NY 14260.
Email: haimonti@buffalo.edu

March 17, 2023

Education

Ph.D. (2007) University of Maryland, Baltimore County.
Computer Science and Electrical Engineering.
Thesis: *Empowering Scientific Discovery by Distributed Data Mining on the Grid Infrastructure*.
Advisor: Professor Hillol Kargupta.
M.S. (2002) Temple University. Computer and Information Science.
B.C.S.E. (1999) Jadavpur University. Computer Science and Engineering.



Employment

2022-present: Associate Professor, Department of Management Science and Systems, The State University of New York, Buffalo, NY.
2014-2022: Assistant Professor, Department of Management Science and Systems, The State University of New York, Buffalo, NY.
2016-Present: Core Faculty Member, Computational and Data-Enabled Science and Engg (CDSE), The State University of New York, Buffalo, NY.
2007-2014: Associate Research Scientist, The Center for Computational Learning Systems, Columbia University, New York.
2012-2016: Affiliated Member, Health Analytics and Foundations of Data Science Center(s), Data Science Institute, Columbia University, New York.
2011-2012: Adjunct Assistant Professor, Department of Computer Science, Columbia University, New York.

University at Buffalo
The State University of New York

Research

Broad Area: Machine Learning, Distributed Optimization, Large-Scale Distributed and Parallel Mining, Operations Research.
Primary Focus: Consensus-based Machine Learning

AI is the "New Electricity"
– Andrew Ng, computer scientist and entrepreneur

> "**Own an iPhone X? Its facial recognition system is powered by AI. Ever been redirected by Google Maps because of an accident or construction ahead? You guessed it: AI.**
> *And those are just a couple of small examples. By one estimate, AI contributed a whopping 2 trillion to global GDP last year. By 2030, it could be as much as $15.7 trillion, "making it the biggest commercial opportunity in today's fast-changing economy," according to a recent report by PwC."*

– Excerpt from Forbes, `https://www.forbes.com/sites/greatspeculations/2019/02/25/ai-will-add-15-trillion-to-the-world-economy-by-2030/?sh=469e74431852`]

Acknowledgements

Chitra Anusandhan: Art Recommendation from Painted Narrative Scrolls

Chitra= Picture; Anusandhan = Search

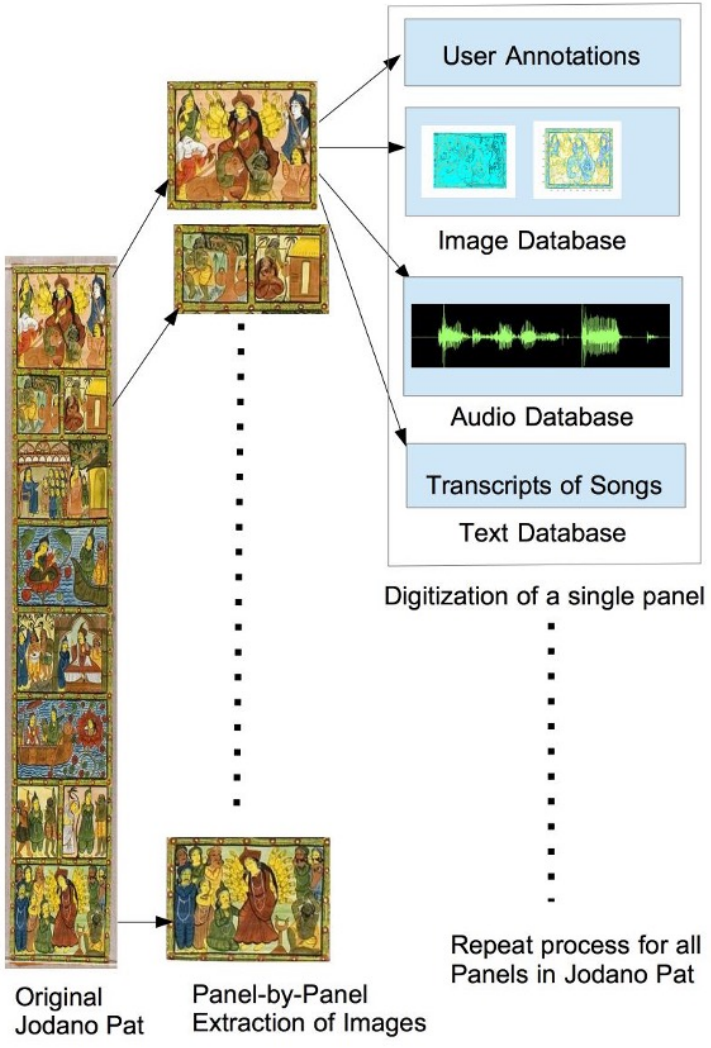
Digital Humanities

- An art conservation project from South Asia
- Collaboration between the South Asian Studies Program, Computer Science and Operations Research
- Example of collaboration between anthropology, art, computer science, and operations research
- Goal: Build a recommendation engine for folk art

Data Collection: An endangered performing art

A video clip obtained in field research:
https://www.youtube.com/watch?v=rJcrD1Nv-Jo

Data - What Kind?



Data - Closeup

Oh Rangila (lit. the coloured one) I am going to arrange the marriage of the fishes today. Oh Rangila, I am going to arrange the marriage of the fishes today. Tangra fish says, I'll be the ear ring Rangila. Oh Rangila, I am going to arrange the marriage of the fishes today. Pankal fish says, I'll be the necklace Rangila.

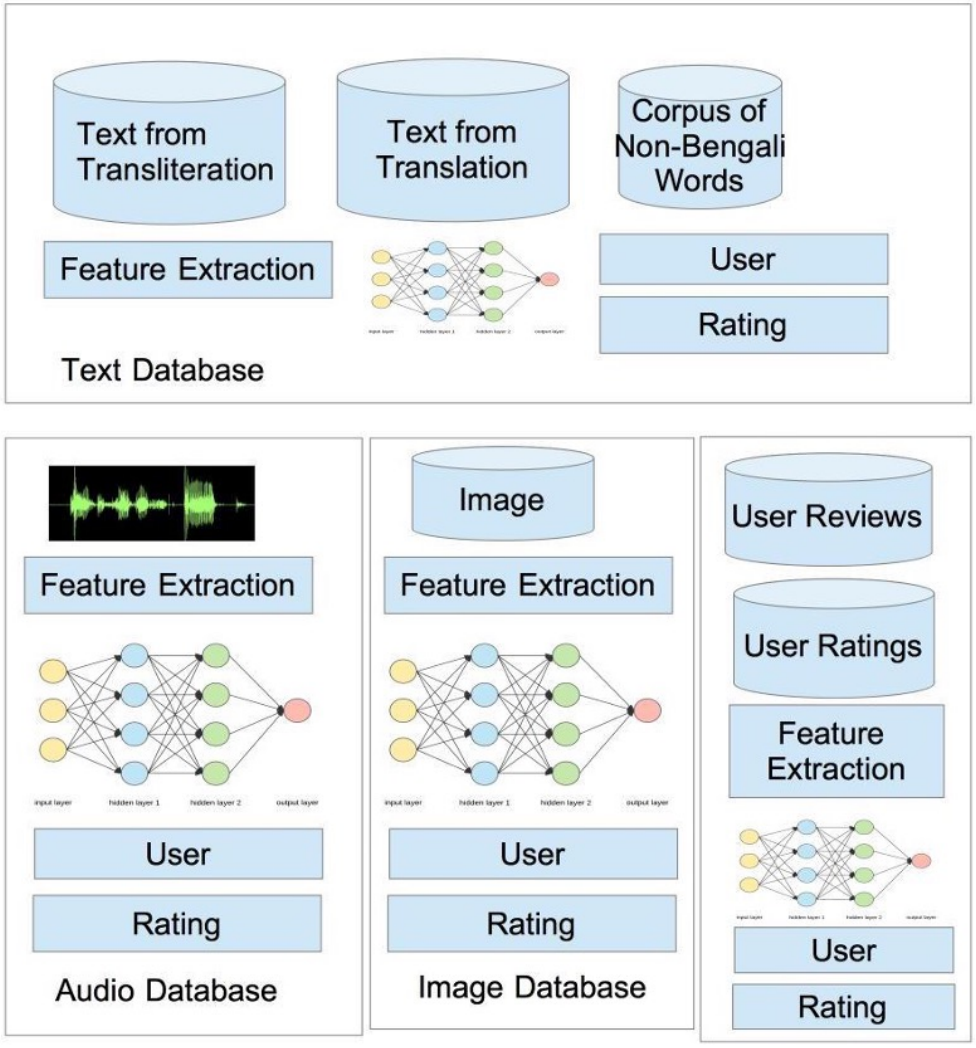
Table 2: Translation of song, obtained from `http://naya.research.wesleyan.edu/the-wedding-of-the-fish-meena/`

A Panel / Frame from the scroll and the associated song to be sung

Multimodal Data

- Different types of sensory information can be collected from our environment – vision, sounds, smell, etc.
- These different *types* are often called different *modalities*
- Multiple modalities help develop a reliable understanding of the world.
- Use multiple modalities to model the world around us.

System Architecture



Learning Problem

Given an image (panel of scroll) and associated lyrics of the song, are there:

- Mythological characters
- Trees
- Animals

in it?

Human Annotation

- Before modeling using statistical techniques, can we have humans identify mythological characters, trees or animals in the scrolls? How good are we at this task?
- Krippendorf's Alpha: Is a reliability coefficient developed to measure the agreement among observers, coders, judges, raters, or measuring instruments drawing distinctions among typically unstructured phenomena or assign computable values to them
- Also known as inter-coder agreement or inter-rater reliability
- Recruit a group of coders and ask them to annotate the concepts

Task	Krippendorf's Alpha
Mythological Figures	0.69
Trees	0.85
Animals	0.69

Feature Extraction

- Pixel-based features, Regions of Interest based features (shapes, moments)
- Very High Dimensional Data
- Use these features and Labels to find out whether we can learn the concepts - mythological features, trees and animals.

Multimodal Data Analysis: Some Results

Animal				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.1659**	**1249.2360**	**26.91%**	**75.24%**
Stdev	0.0109	1795.6609	3.90%	8.51%

Tree				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.1710**	**1362.7319**	**15.48%**	**76.19%**
Stdev	0.0080	2930.5283	3.00%	6.73%

Myth 1				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.1985**	**839.4618**	**18.60%**	**56.19%**
Stdev	0.0022	1128.6770	1.29%	6.21%

Myth 2				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.1344**	**289.4307**	**29.06%**	**82.86%**
Stdev	0.0101	542.7780	4.89%	7.22%

Figure: Results using only features generated from images

Multimodal Data Analysis: Some Results

Animal				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0991**	**18.9306**	**56.29%**	**65.71%**
Stdev	0.0072	17.3450	3.43%	8.52%

Tree				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0879**	**12.0430**	**56.56%**	**74.28%**
Stdev	0.0069	11.2557	3.32%	15.28%

Myth 1				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.1050**	**7.4241**	**56.91%**	**64.76%**
Stdev	0.0118	5.5178	5.31%	5.43%

Myth 2				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0778**	**6.8188**	**58.95%**	**75.24%**
Stdev	0.0123	3.9580	6.34%	11.37%

Figure: Results using only features generated from text

Multimodal Data Analysis: Some Results

Animal				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0574**	**2794.7566**	**74.63%**	**67.62%**
Stdev	0.0061	4106.4331	3.48%	14.05%

Tree				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0609**	**4468.1775**	**69.78%**	**66.67%**
Stdev	0.0133	8839.4870	7.02%	13.89%

Myth 1				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0684**	**717.9587**	**71.92%**	**57.14%**
Stdev	0.0089	1230.7645	3.95%	5.83%

Myth 2				
	Test Error	Valid Error	R Square	Accuracy
Mean	**0.0514**	**253.1198**	**72.87%**	**71.43%**
Stdev	0.0082	323.9048	4.13%	9.52%

Figure: Results using only features generated from Image and Text

Take Away Message from Initial Results

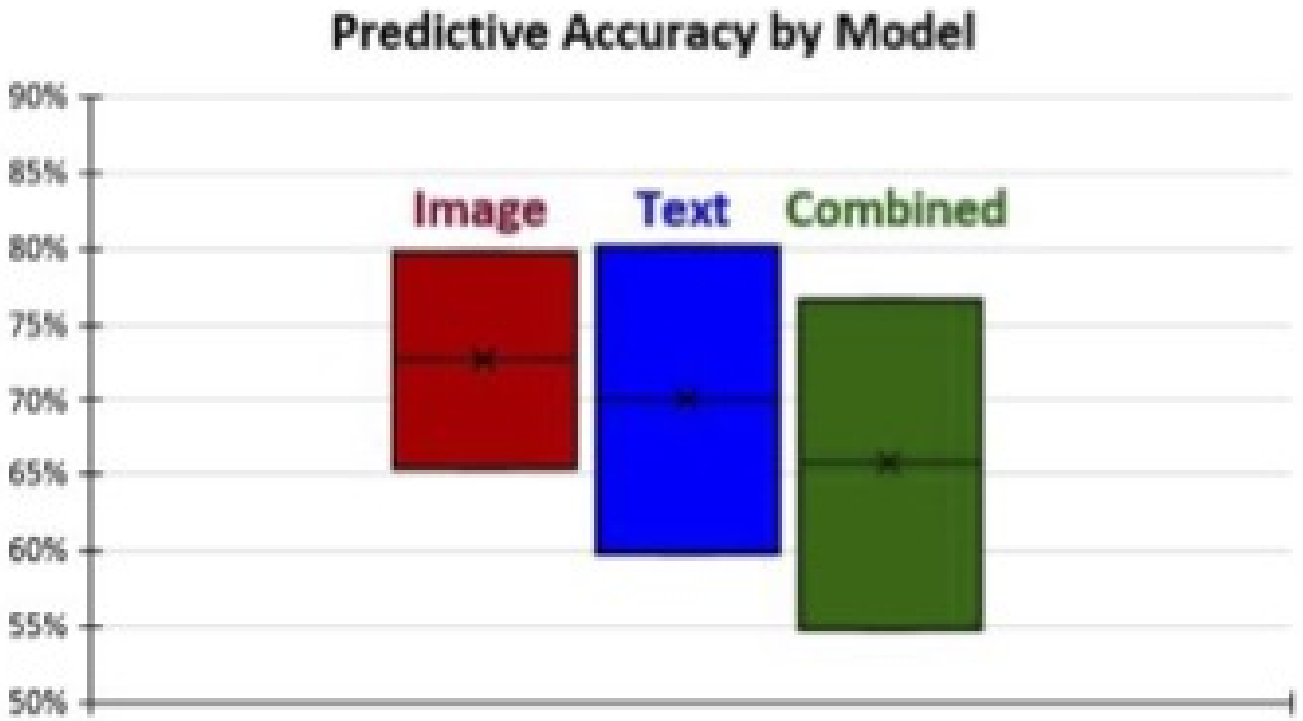


Figure: Results using only features generated from Images, Text, and Image and Text

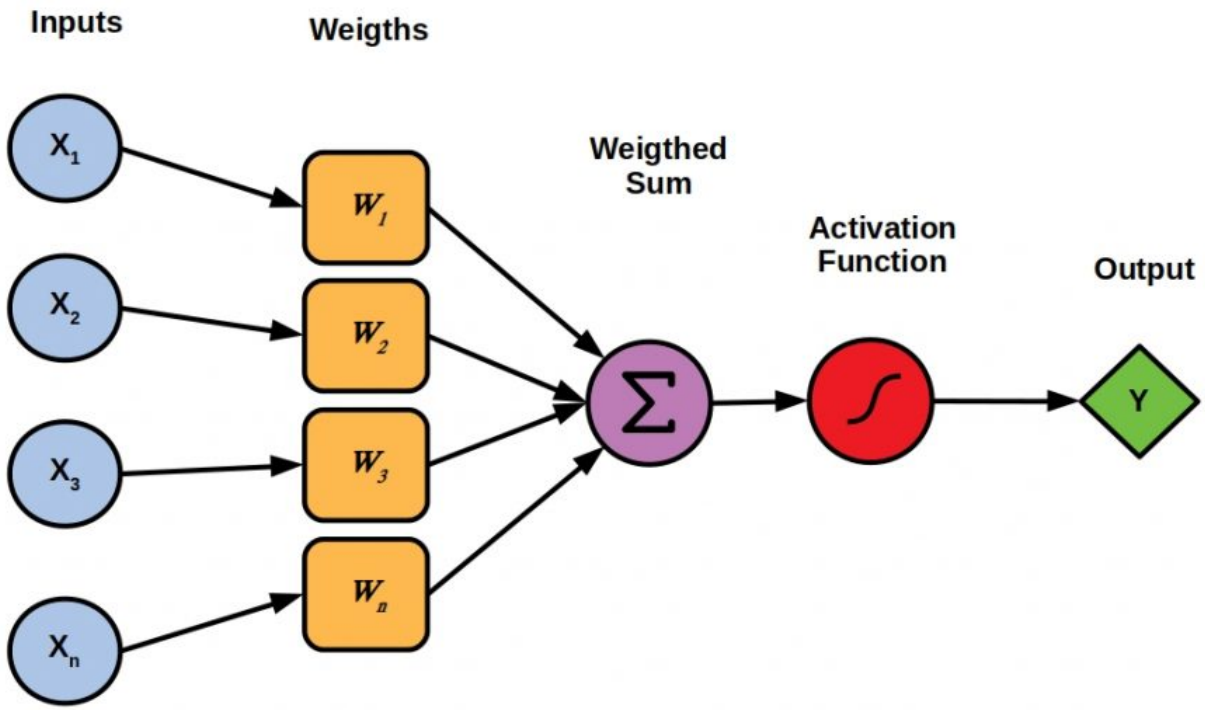
Challenges in dealing with multimodal data

- Obtain a compact and modality invariant representation – shared representation
- Translate data between modalities – cross-modal generation
- Dealing with heterogeneous data modalities

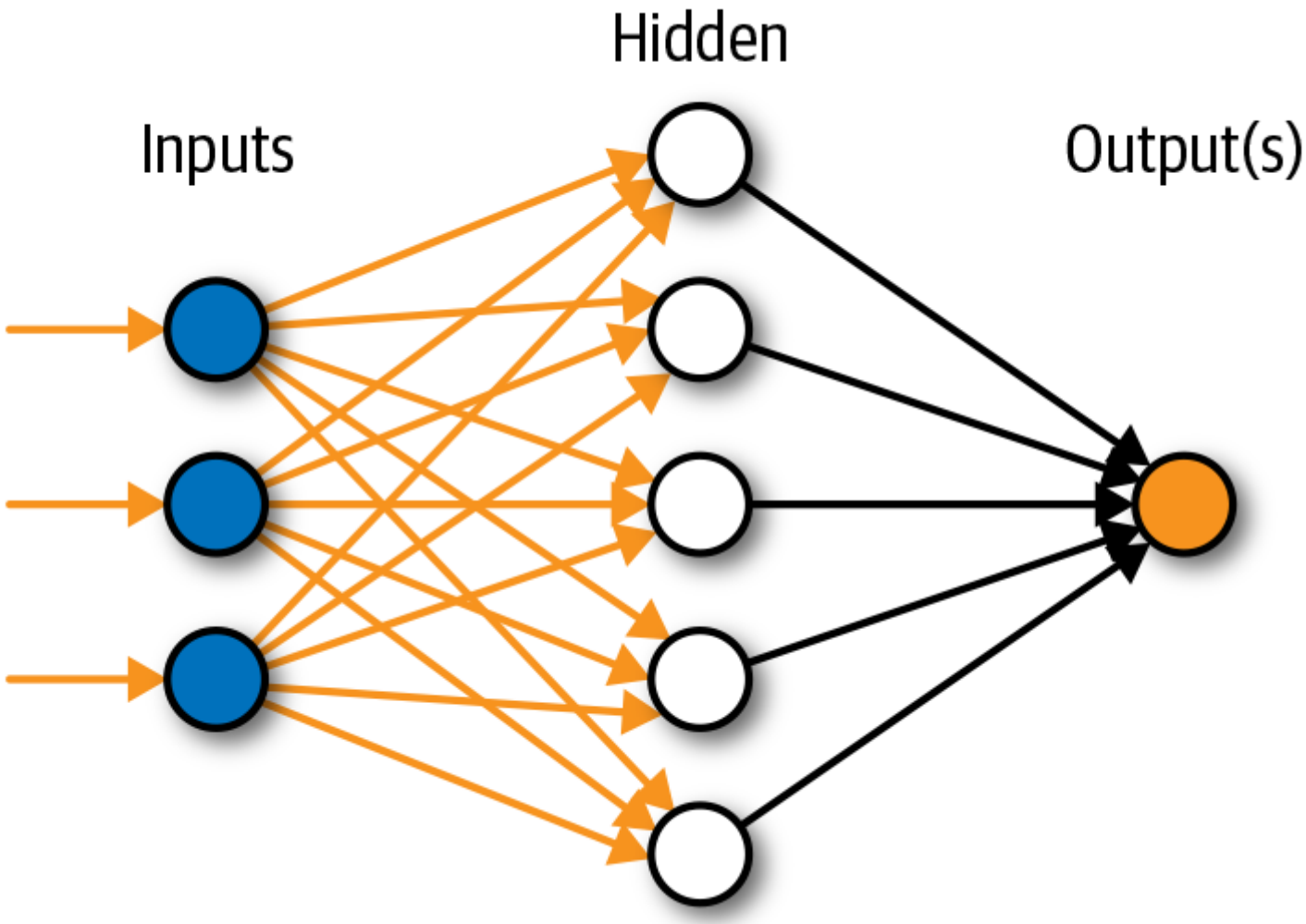
What are Perceptrons?

- A type of artificial neural network
- Takes a vector of real-valued inputs and calculates a linear combination of these inputs
- Outputs a 1, if the results is greater than some threshold or -1 otherwise
- Perceptron represents a hyperplane decision surface in n-dimensional space of instances (data points)

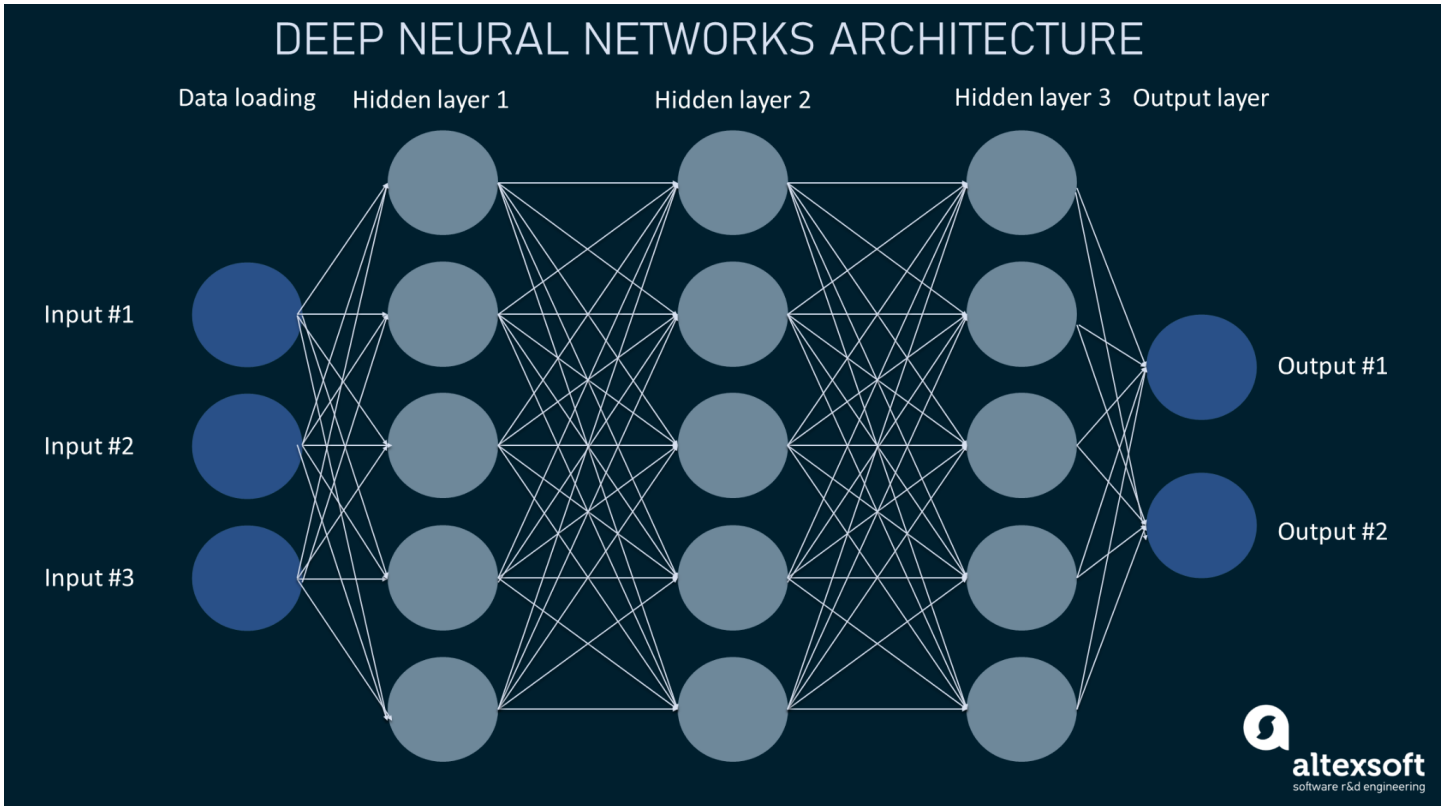
Deep Learning 101: Perceptrons



What are Artificial Neural Networks?



What are Deep Neural Networks?



Generative Models

What are they?

Goal: Treat all multimodal data as stochastically generated from a global latent variable that corresponds to a shared representation.
Advantages:

- Not only multimodal data can be generated from a shared latent variable, but also the latent variable can be inferred from arbitrary modalities.
- Helps acquire shared modality and cross-modal generation.

Disadvantages: Very high dimensional feature space, compute intensive.

Deep Generative Models

What are they?

- Framework that represents the distribution of generative models by *deep* neural networks.
- Characterized by end-to-end learning with back-propagation
- Ability to generate high-dimensional and complex data

Some examples of Deep Generative Models

- Variational Autoendcoders (VAEs)
- Generative Adversarial Networks (GANs).

Multimodal Generative Models

- An i.i.d dataset $X = \{X^{(i)}\}_{i=1}^{N}$
- Each example $X^{(i)}$ has M modalities
- True joint distribution of multimodal data: $p_{data}(X) = p_{data}(x_1, \cdots, , x_M)$
- Assume $z^{(i)} \in Z$ embeds different modalities $X^{(i)}$

Problem Setting

- Shared Representation: Embed all modalities in X in a good common space
- Cross Modal Generation: Generate modalities, when only some of the M modalities are known using the shared representation.

What does it mean to have a good shared representation?

- Similarity in the representation space must imply similarity of the corresponding "concepts"
- Must be easy to obtain even in the absence of some modalities.

Multimodal Generative Models

What are they?

Consist of Latent Variable models with all modalities X as observed variables and the shared representation z as a latent variable. Each modality is assumed to be conditionally independent given a latent variable.

$$p_\theta(X, z) = \prod_{m=1}^{M} p_{\theta_m}(x_m|z)p(z) \tag{1}$$

where $\theta = \{\theta_m\}_{m=1}^{M}$ is the set of parameters for the conditional distributions of each modality.

Multimodal Generative Models

Maximize Loglikelihood

$$\hat{\theta} = arg \max_{\theta} E_{p_{data}(x)}[\log p_\theta(X)] \tag{2}$$

- Objective is intractable, and involves marginalization of a latent variable
- Alternatively, infer from the generative model i,e, find the posterior distribution $p_\theta(z|X_k)$ of the shared representation z given any modality x_k
- Even estimating posterior is intractable!

Key issues in multimodal generative learning

- How to design and train generative models
- How to perform inference from a latent variable

What are Variational Auto Encoders (VAEs)?

- They are latent variable models
- $p_\theta(x) = \int p_\theta(x|z)p(z)dz$ where $p_\theta(x|z)$ is parameterized by a deep neural network and prior $p(z)$ is often taken to be a standard multivariate Gaussian.
- Instead of maximizing the intractable log-likelihood, we maximize an Evidence Lower Bound (ELBO)

Advantages of Variational Auto Encoders (VAEs)?

- VAEs represent both generation and inference as paths of DNNs – can be trained fast and handle high dimensional data
- Good models for representation learning
- Can deal with heterogeneity of multimodal data

Two categories of multimodal deep generative models

Two main categories based on how they model inference from shared representations

- Coordinated Models: modeling inference from a single modality
- Joint: modeling inference from ALL modalities.

A little more on Coordinated and Joint Models

- Coordinated model can embed each modality into the same shared representation; cannot perform inference on all modalities
- Joint model directly models inference to the shared latent space given all modalities

Known Challenges for Joint Variational Encoders

- Handling of missing modalities
- Modality specific latent variables
- Weakly supervised learning

Types of Deep Generative Models

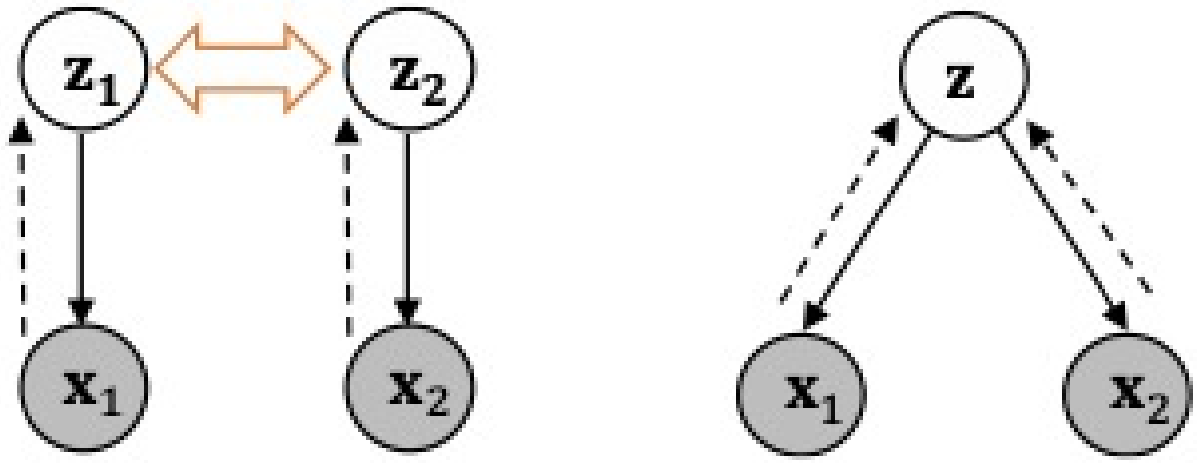


Figure: Coordinated and Joint Deep Generative Models

Future Research Plan

- Can work for two or three modalities; in real-world number of modalities can be very large!
- Greater number and modality of data can cause difficulty in training them.

Work in Progress

- We are always looking for experienced programmers (Java, Python) to help our efforts!
- Coursework / Independent study options or opportunities to participate hands on in our projects.
- Work with a team of like-minded business analytics

Support our artists!

Our artists

We are always on the look out for sponsors for our artists!

- Buy art-work from collaborators
- Support their work through exhibitions
- Help preserve their work in museums and other indexed digital archives to make them searchable.