Deep Causality Variational Autoencoder Network for Identifying the Potential Biomarkers of Brain Disorders

Amani Alfakih, Zhengwang Xia, Bahzar Ali, Saqib Mamoon, and Jianfeng Lu

01.24.25
Joanne Wardell - MLBBQ

COMPANY NAME

Generative AI

- What is it?
- How does it work?
- What are its uses in neuroimaging, causality, modeling, etc?
- Causal VAE paper

Images generated using stable diffusion

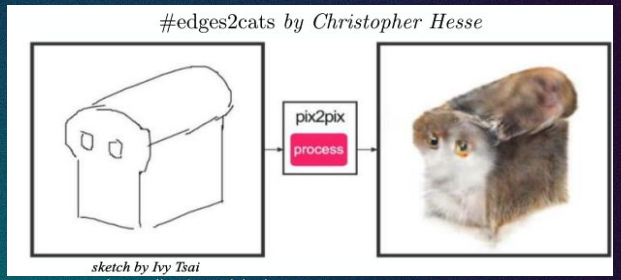
https://civitai.com/models/15003

Why Generative AI

Many right answers

Caption to image generation

https://openreview.net/pdf?id=Hyvw0L9el

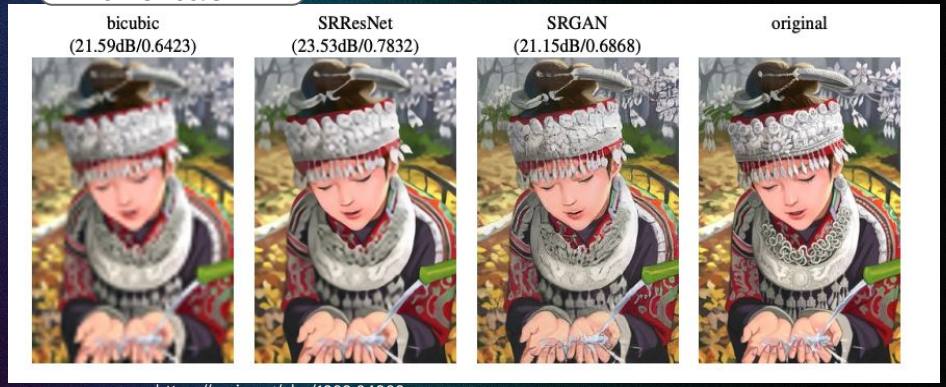


https://arxiv.org/abs/1611.07004

Sketch to image generation

Intrinsic to the task

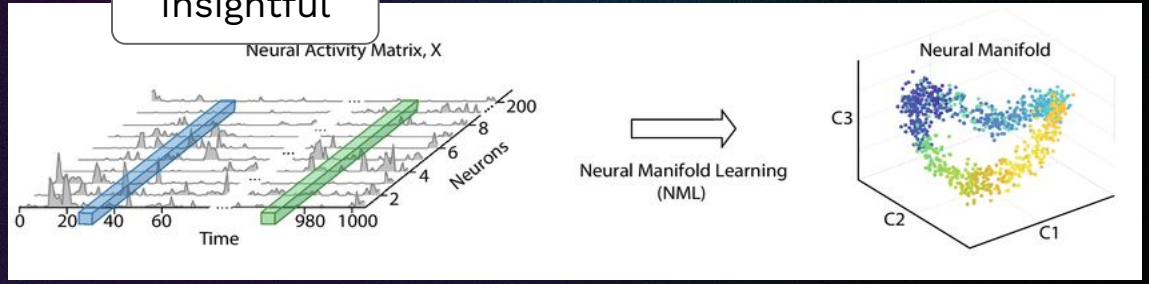
Image super resolution



https://arxiv.org/abs/1609.04802

Insightful

Neural manifold learning



Dimensionality reduction and generative models

To appear as a part of an upcoming textbook on dimensionality reduction and manifold learning.

Factor Analysis, Probabilistic Principal Component Analysis, Variational Inference, and Variational Autoencoder: Tutorial and Survey

Benyamin Ghojogh BGHOJOGH@UWATERLOO.CA
Department of Electrical and Computer Engineering,
Machine Learning Laboratory, University of Waterloo, Waterloo, ON, Canada

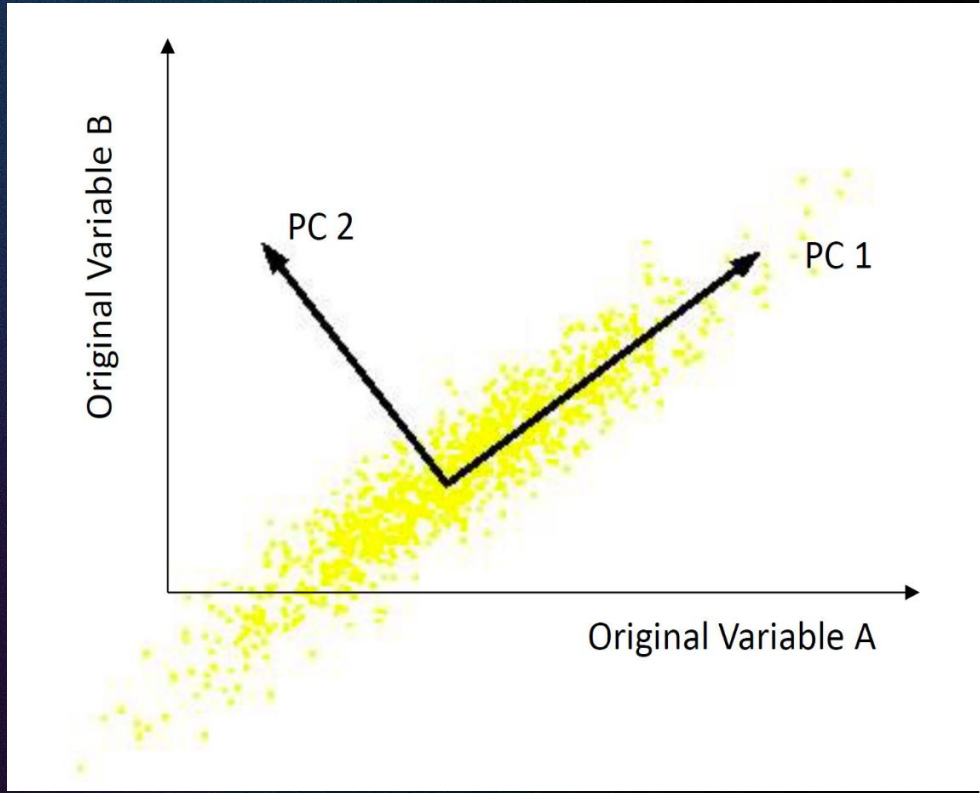
Ali Ghodsi ALI.GHODSI@UWATERLOO.CA
Department of Statistics and Actuarial Science & David R. Cheriton School of Computer Science,
Data Analytics Laboratory, University of Waterloo, Waterloo, ON, Canada

Fakhri Karray KARRAY@UWATERLOO.CA
Department of Electrical and Computer Engineering,
Centre for Pattern Analysis and Machine Intelligence, University of Waterloo, Waterloo, ON, Canada

Mark Crowley MCROWLEY@UWATERLOO.CA
Department of Electrical and Computer Engineering,
Machine Learning Laboratory, University of Waterloo, Waterloo, ON, Canada

Principal Components (PCs)

Principal components are orthogonal directions of greatest variance in data. When the data is projected onto the PC1 axis it shows the greatest “spread”.



Principal Component Analysis (PCA)

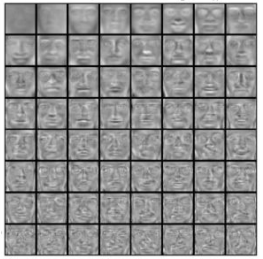
Principal component analysis

- data reduction approach
- seeks to find optimal linear transformation
  - preserve as much information in the data as possible
  - projects the data onto its principal components

Olivetti Faces Dataset

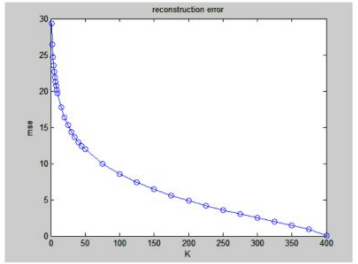
400 face images
40 people
10 images per person

Top eigenvectors



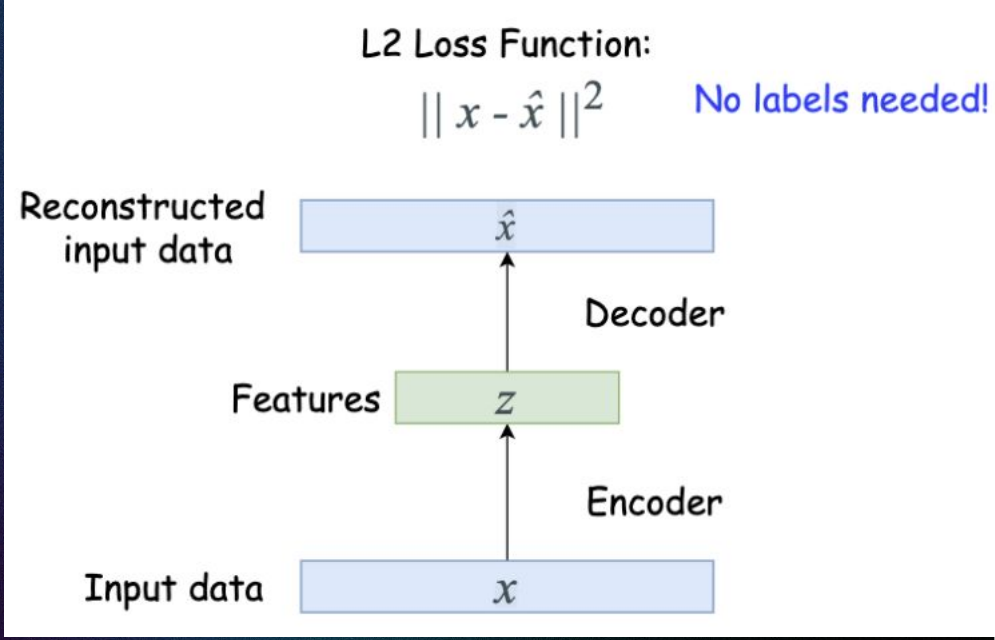
Average face

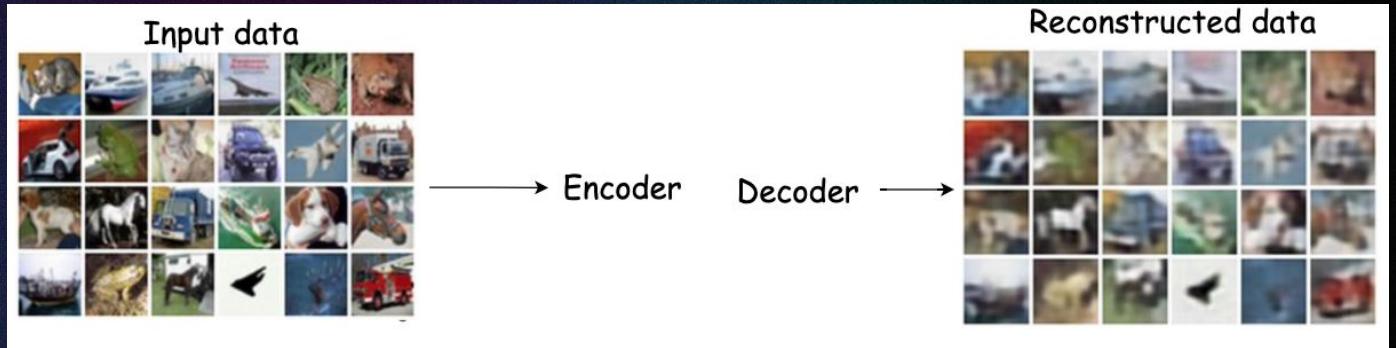
Face Reconstruction



Autoencoder (AE)

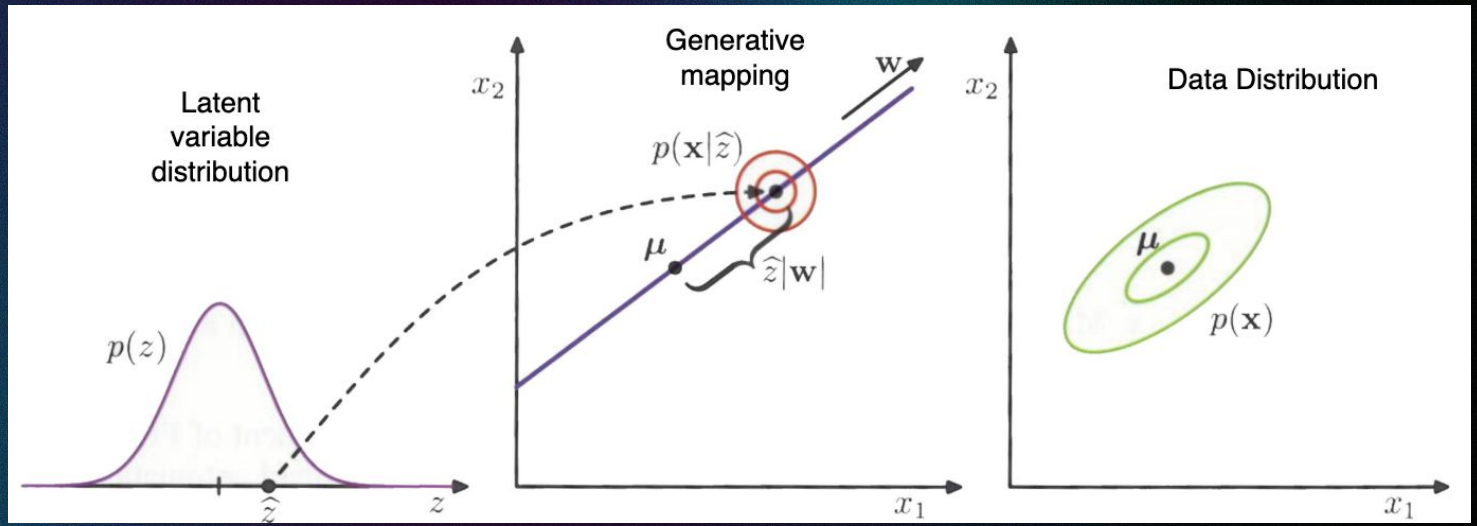
An **autoencoder** is an artificial neural network that learns efficient encodings of data. AEs are like a non-linear form of PCA.





Factor Analysis (FA)

Factor analysis is a generative model that assumes the data we observe is generated from a set of latent variables, assumes a linear relationship between observed and latent variables.

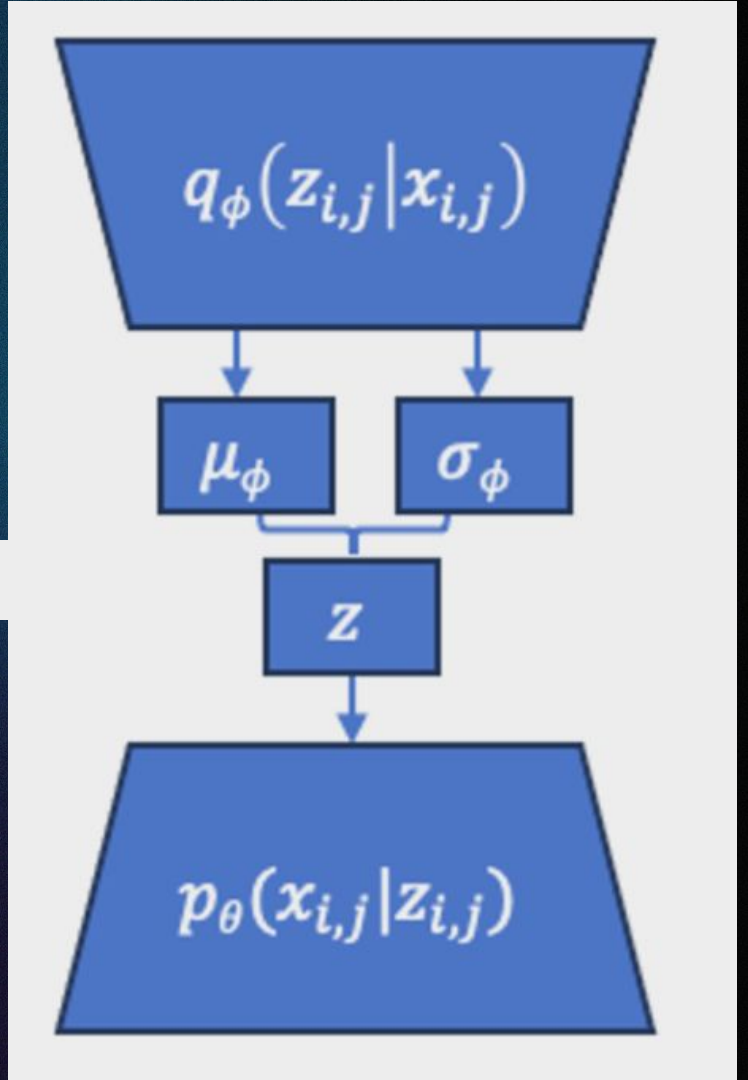


Bishop – Pattern Recognition and Machine Learning

Variational Autoencoder

A **variational autoencoder** uses probabilistic encoders/decoders rather than deterministic ones. It learns a latent space from which new data can be sampled.

Can be thought of as a non-linear form of factor analysis.



Need Objective function to train the VAE

Factor analysis assumes the data we observe is generated from a set of latent variables. Let X be a random variable representing the observed data and let Z be a random variable representing the set of latent variables.

- $p(X)$
  - Maximize the likelihood of training data
- $p(Z \mid X)$
  - Model the latent factors given the observed data

Intractability

- Consider the conditional probability: $p(Z \mid X)$
- Use Bayes' Rule: $p(Z \mid X) = \frac{p(X \mid Z)p(Z)}{p(X)}$
- How can we evaluate $p(X)$?
- Since $p(X)$ consists of a number of latent variables, $Z=z_1, \ldots, z_n$, then the evaluation would be

$$\int_{z_n} \cdots \int_{z_1} p(Z)p(X \mid Z)\, dz_1 \ldots dz_n$$

  - .....typically intractable due to the combinatorial complexity of Z

Variational Inference

- Use variational inference, a technique based on the calculus of variations, to estimate the intractable probability $p(Z \mid X)$.

1. Find an estimator $q(Z)$
2. to match $p(Z \mid X)$
3. s.t. $q(Z) \approx p(Z \mid X)$

Use the Kullback–Leibler divergence as a measure of fitness of $q(z)$.

$$\mathrm{KL}\left[q(Z) \,\|\, p(Z \mid X)\right]$$

Evidence Lower Bound

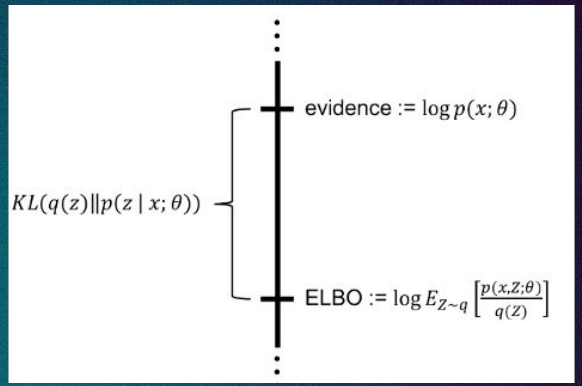
$$= \mathbb{E}_{Z\sim q(Z)} \log \left[\frac{q(Z)}{p(Z \mid X)} \right] \quad \text{(definition of KL divergence)}$$

$$= \mathbb{E}_{Z\sim q(Z)} \log[q(Z)] - \mathbb{E}_{Z\sim q(Z)} \log[p(Z \mid X)] \quad \text{(logarithms)}$$

$$= \mathbb{E}_{Z\sim q(Z)} \log q(Z) - \mathbb{E}_{Z\sim q(Z)} \log \left[\frac{p(Z, X)}{p(X)} \right] \quad \text{(Bayes' Theorem)}$$

$$= \mathbb{E}_{Z\sim q(Z)} \log q(Z) - \mathbb{E}_{Z\sim q(Z)} \log p(Z, X) + \mathbb{E}_{Z\sim q(Z)} \log p(X) \quad \text{(logarithms)}$$

$$= \mathbb{E}_{Z\sim q(Z)} \log p(X) - \mathbb{E}_{Z\sim q(Z)} \log \left[\frac{p(Z, X)}{q(Z)} \right] \quad \text{(regrouping)}$$



$$KL\left[q(Z) \,||\, p(Z \mid X)\right] = \log p(X) - \mathbb{E}_{Z\sim q(Z)} \log \left[\frac{p(Z, X)}{q(Z)} \right]$$

'fitness' of q(Z)

evidence

evidence lower bound

ELBO and the VAE

$$\mathbb{E}_{Z\sim q(Z)} \log \left[\frac{p(Z,X)}{q(Z)}\right] \quad \text{(ELBO)}$$

$$= \mathbb{E}_{Z\sim q(Z)} \log \left[\frac{p(X \mid Z)p(Z)}{q(Z)}\right] \quad \text{(Bayes' Theorem)}$$

$$= \mathbb{E}_{Z\sim q(Z)} \log p(X \mid Z) + \mathbb{E}_{Z\sim q(Z)} \log \left[\frac{p(Z)}{q(Z)}\right] \quad \text{(Logarithms)}$$

For the VAE, the inference network is approximated as $q_\phi(z \mid x)$ and the generative network is $p_\theta(x \mid z)$.

$$= \mathbb{E}_{Z\sim q_\phi(z|x)} \log p_\theta(x \mid z) + \mathbb{E}_{Z\sim q_\phi(z|x)} \log \left[\frac{p_\theta(z)}{q_\phi(z \mid x)}\right]$$

$$= \underbrace{\mathbb{E}_{Z\sim q_\phi(z|x)} \log p_\theta(x \mid z)}_{\text{Reconstruction loss}} - \underbrace{KL\left[q_\phi(z \mid x) \,||\, p_\theta(z)\right]}_{\text{Make appx. posterior close to prior}}$$

The structure of the unobserved

Structural equation models (SEMs) - the problem of determining the causal structure among a set of unmeasured variables of interest.

General form:

$$x_i = f_i(\mathrm{pa}_i, u_i), \quad i = 1, \ldots, n$$

where pa_i (connoting parents) stands for the set of variables judged to be immediate causes of X_i and U_i represents errors or disturbances due to omitted factors.

Linear SEMs

Linear form:

$$x_i = \sum_{k \neq i} \alpha_{ik} x_k + u_i, \quad i = 1, \ldots, n$$

In the linear form, pa_i corresponds to the variables on the r.h.s. that have non-zero coefficients.

x_i - dependent variable we are modeling
x_k - independent variables that influence x_i
α_{ik} - coefficients rep. strength & direction of relationships b/w vars
u_i - error term, accounting for unobserved influences / noise

Also Known As

SEMs may also be referred to as:

- Covariance structure analysis
- Analysis of moment structures
- Analysis of linear structural relationships
- Causal modeling

What are latent variables?

- Social scientific concepts which are not directly observable (makes them hypothetical, or, latent)
  - Intelligence
  - Social capital

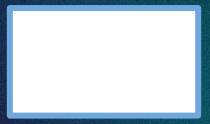
- Measurable using observable indicators

- For example, the variance of a questionnaire item as being caused by
  - latent variable (which we would like to measure)
  - Other factors (like error/ unique variance)

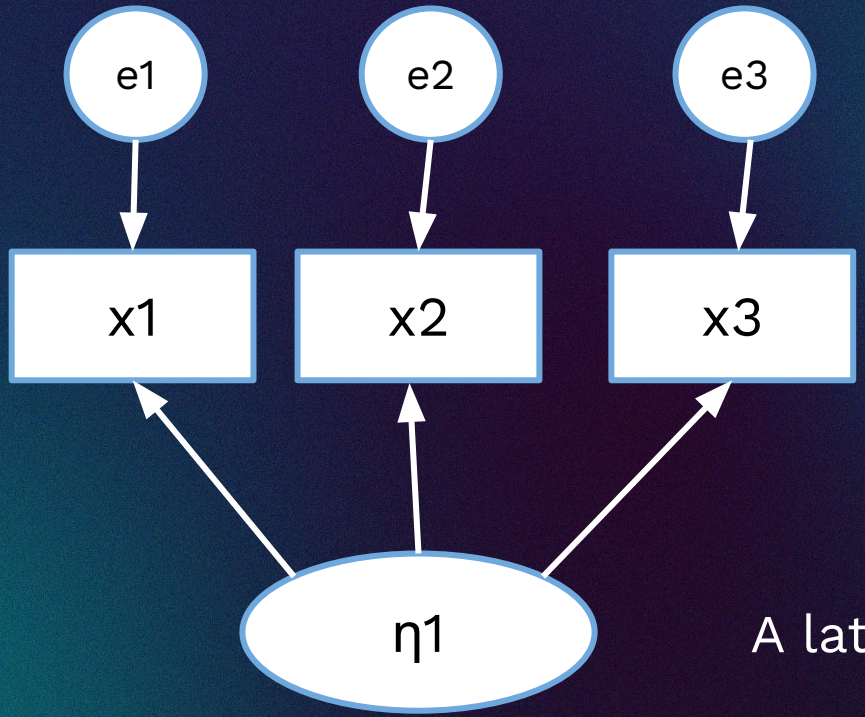
Exogenous/Endogeneous Variables

- Endogenous (dependent)
  - Caused by variables in the system
- Exogeneous (independent)
  - Caused by variables outside the system
- In SEM a variable can be a predictor and an outcome (a mediating variable)

Path Diagram Conventions

- Measured latent variable 
- Observed / manifest variable 
- Error variance / disturbance term 
- Covariance / non-directional path 
- Regression / directional path 

Path Diagram Example

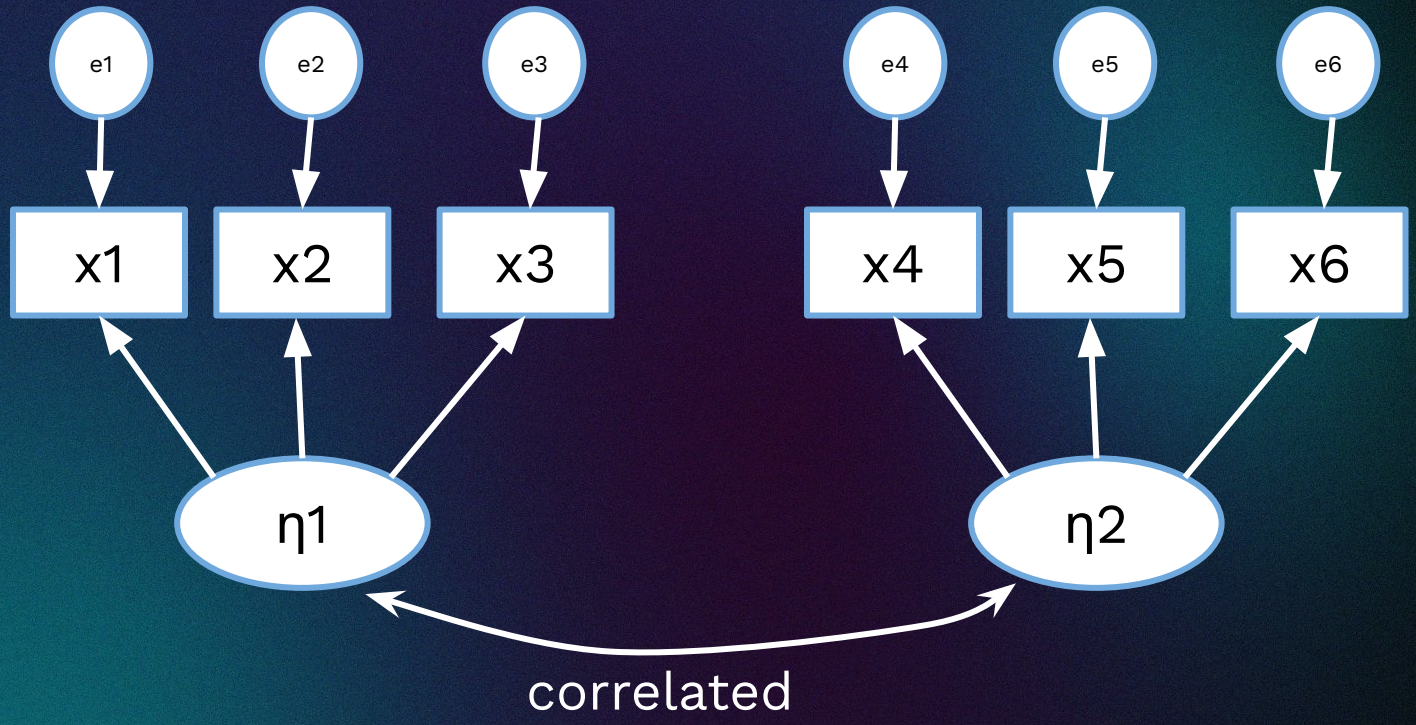


With 3 error variances

Causes/measured by
3 observed variables

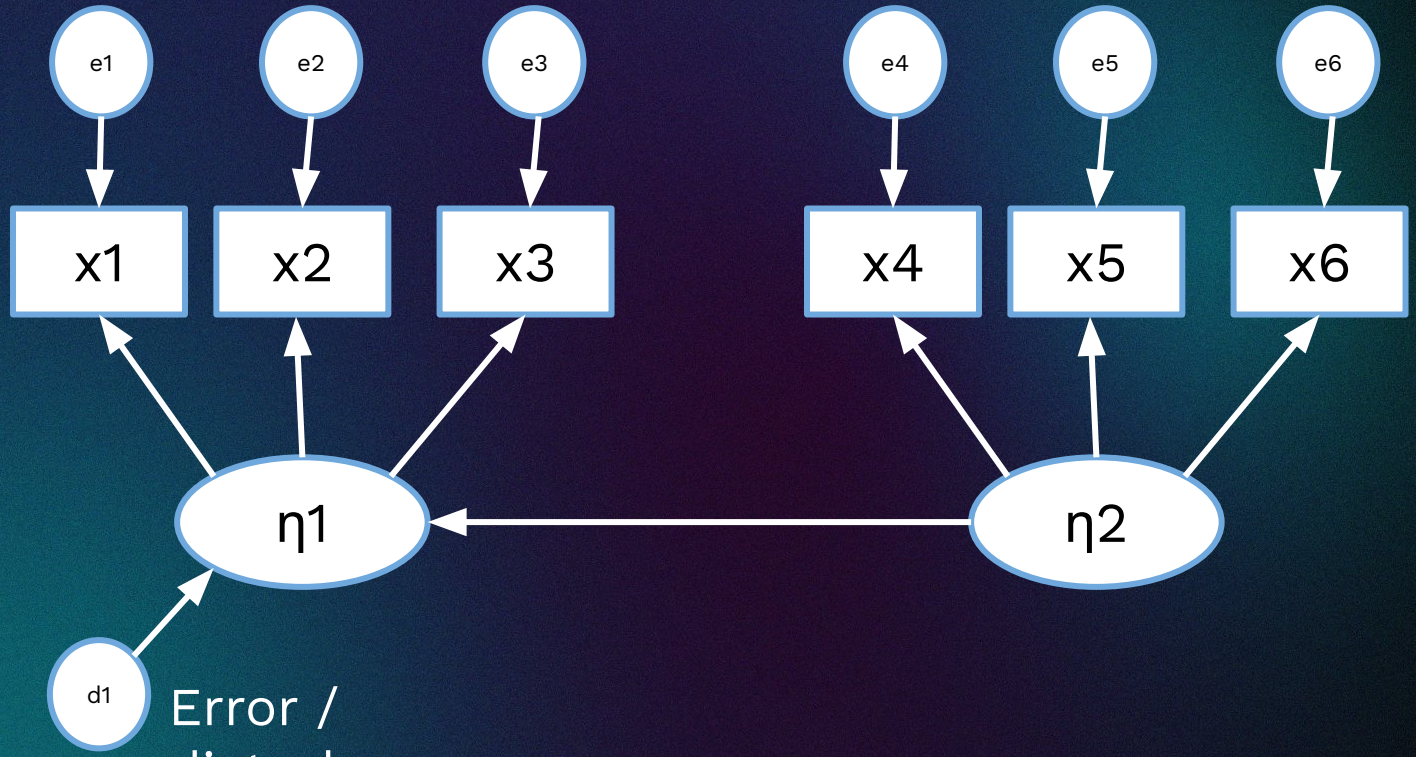
A latent variable

Path Diagram Example



2 latent variables, each measured by 3 observed variables

Path Diagram Example



Error / disturbance

2 latent variables, each measured by 3 observed variables

Paper - Problem Setup

- Suppose $A \in \mathbb{R}^{m \times m}$ is a weighted graph with m nodes.
- According to the SEM theory, the causal relationship between m brain regions can be expressed by $X = A^T X + Y = M^{-1} Y$.
- $Y \in \mathbb{R}^{m \times d}$ - gaussian exogenous factors
- $X \in \mathbb{R}^{m \times d}$ - fMRI time series data
- d - length of time series
- M - structural relationship between brain regions.
- A - directed influence of regions on each other.

Paper - Problem Statement

- Given the distribution of Y and samples $X_1, \ldots, X_n$
  - X - sample index
  - n - total number of training samples
- Use the VAE to reconstruct a sample X_k, $k \in \{1,\ldots,n\}$.

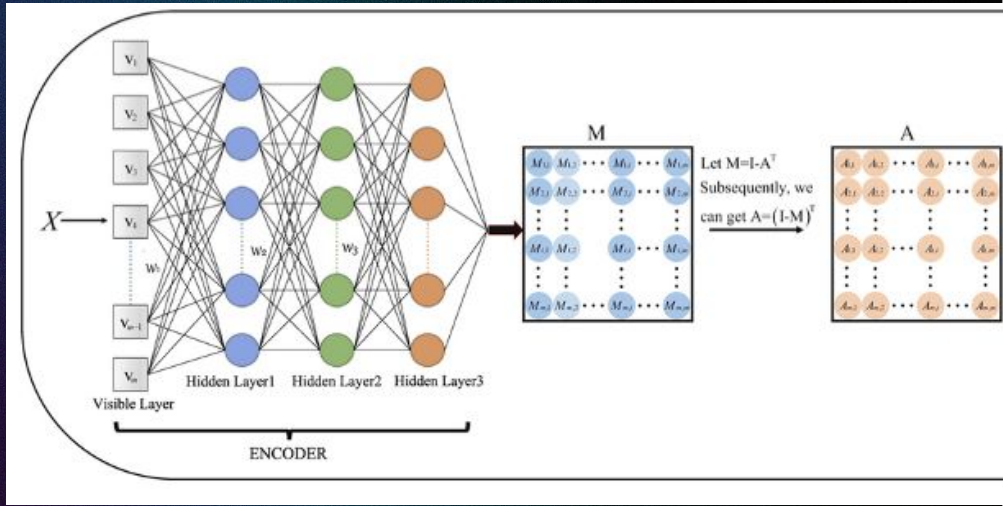
- Subject to maximizing the evidence lower bound:

$$\mathrm{ELBO}_k = \mathbb{E}_{q(Y|X_k)}\left[\log p(X_k \mid Y)\right] - \mathrm{KL}\left[q(Y \mid X_k) \| p(Y)\right]$$

Encoder Architecture

Time series data X input to the encoder

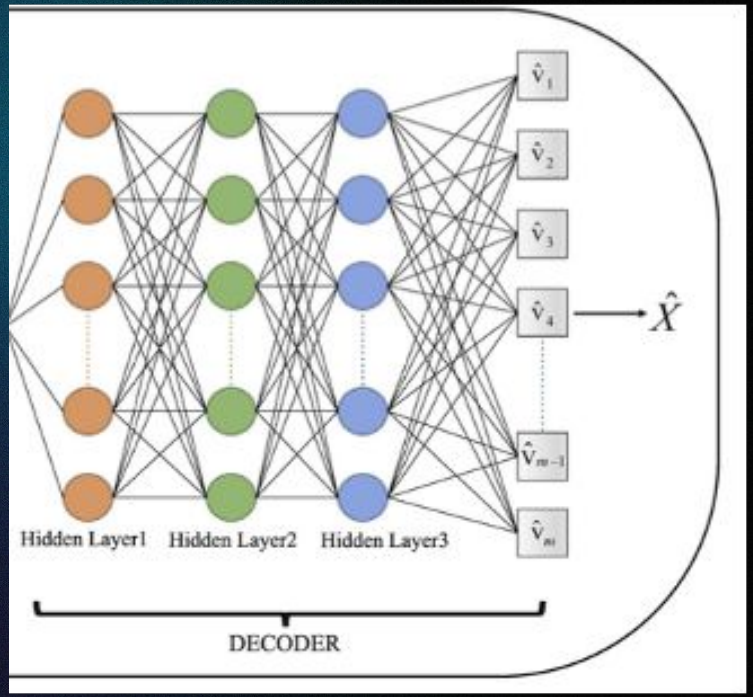
- An MLP learns parameters M_Y and $\log_S Y$
- These are multiplied with M
- Y is sampled from a Gaussian
  - with mean M_Y and std. dev. S_Y
- Then Y is multiplied with M^{-1}
- A is solved as $A=I-M^T$



Decoder Architecture

An MLP takes $M^{-1}Y$ as input

- Learns parameters M_X and $\log_S X$
- These are used to sample X
  - From Gaussian with
    - Mean M_X, std. dev. S_Y



Experimental Procedure

- The VAE is first trained using the ELBO objective.
- Next, the encoder is used to embed the entire dataset to its causal graph representation.
- Causal graph weights are used as raw features and the recursive feature elimination (RFE) is used as a feature selector.
- These are used as an input vector to a ridge classifier which is trained to predict the group label from the feature vector.

Experimental Procedure

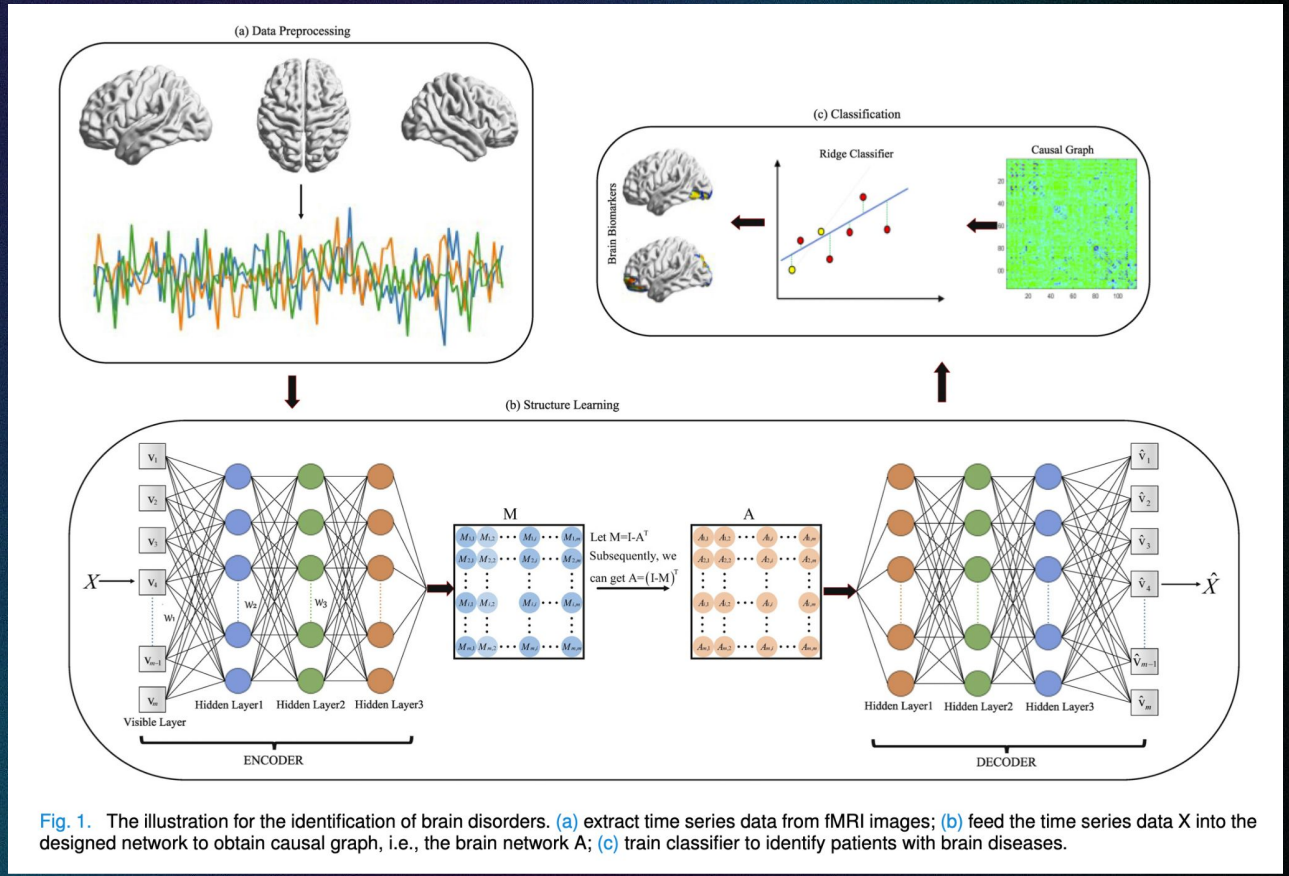


Fig. 1. The illustration for the identification of brain disorders. (a) extract time series data from fMRI images; (b) feed the time series data X into the designed network to obtain causal graph, i.e., the brain network A; (c) train classifier to identify patients with brain diseases.

Datasets

- 2 datasets were used: Alzheimer's Disease Neuroimaging Initiative (ADNI) and Autism Brain Imaging Data Exchange (ABIDE)

- Each dataset processed using AAL Atlas parcellation
  - 90 anatomical brain regions
  - Each has a temporal activity record (BOLD signal)

Dataset Name	Participant Count	Group Types	Scan Types
ADNI	214	NC, eMCI, LMCI	fMRI, sMRI
ABIDE	1112	HC, ASD	fMRI, sMRI

Compared Methods

Compared to popular brain modeling methods:

1. Pearson correlation-based method (PC)
    a. Computes linear correlation coefficients
2. Sparse low-rank representation method (SLR)
    a. Decomposes data into pattern matrices
3. Granger causality-based method (GC)
    a. Assesses time series causality
4. Transfer entropy (TE)
    a. Measures time series dependence
5. Linear non-Gaussian acyclic model (LiNGAM)
    a. Infers linear causal relationships
6. Temporal lag neural network (ETLN)
    a. Models temporal dependencies in data
7. Causal recurrent variational autoencoder (CR-VAE)
    a. Learns causal structure from data

Results - ADNI Dataset

- Five runs of each method were performed and the average was used as the final result, with best scores bolded.
-
- Their method Deep CVAE achieves the best classification performance in all tasks with accuracies of 75.6%, 82.6%, and 74.4%.
-

This indicates that their method is learning meaningful causal relationships

TABLE III
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON ADNI DATABASE

Data	Methods	ACC	SEN	SPE	F1
NC& eMCI	PC	72.2±0.6	**70.4±1.1**	73.8±1.5	70.2±0.6
	SLR	66.1±1.7	61.5±2.6	70.1±1.6	62.8±2.1
	GC	65.0±3.0	53.7±4.5	74.8±1.8	58.8±4.1
	TE	62.5±1.0	58.2±2.1	66.2±2.3	59.1±1.2
	LiNGAM	63.1±1.6	59.1±3.8	66.5±3.8	59.8±2.1
	DAG-GNN	64.7±1.3	49.0±2.6	78.4±1.6	56.3±2.0
	ETLN	73.3±1.1	67.2±1.9	78.7±0.6	70.1±1.5
	CR-VAE	62.6±3.3	56.1±3.8	68.3±3.0	58.3±3.8
	Deep CVAE	**75.6±1.5**	69.2±2.1	**81.3±2.3**	**72.4±1.6**
NC& LMCI	PC	78.8±0.8	78.8±2.4	78.9±2.8	78.4±0.8
	SLR	69.1±1.6	68.1±1.2	70.0±2.7	68.3±1.4
	GC	72.7±0.7	63.3±2.8	81.7±2.1	69.4±1.4
	TE	61.8±1.9	60.6±1.5	62.9±3.0	60.8±1.6
	LiNGAM	65.5±1.6	58.8±0.7	72.0±3.5	62.6±1.0
	DAG-GNN	57.1±1.3	49.9±1.5	64.0±1.4	53.2±1.4
	ETLN	78.1±0.9	73.4±2.4	82.6±1.9	76.6±1.2
	CR-VAE	56.4±3.2	50.1±2.6	62.3±4.1	52.9±3.1
	Deep CVAE	**82.6±2.4**	**82.1±2.5**	**83.1±2.9**	**82.2±2.4**
eMCI & LMCI	PC	67.5±1.3	67.8±2.5	67.1±1.3	68.6±1.6
	SLR	65.7±1.6	68.6±2.9	62.6±2.8	67.7±1.8
	GC	61.5±1.4	65.7±1.9	56.9±2.8	64.1±1.3
	TE	52.0±3.1	56.1±4.5	47.4±2.3	55.0±3.5
	LiNGAM	62.3±1.5	69.4±1.3	54.6±2.3	65.8±1.3
	DAG-GNN	53.5±1.2	63.1±4.5	42.9±3.6	58.6±2.2
	ETLN	73.5±2.5	77.4±3.5	69.1±2.9	75.3±2.5
	CR-VAE	58.0±1.3	56.0±3.8	55.7±2.7	59.9±2.2
	Deep CVAE	**74.4±1.7**	**78.7±3.9**	**69.7±2.3**	**76.3±2.1**

Results - ABIDE Dataset

- Deep CVAE performs better than competing methods on most tasks.
- Deep CVAE yields higher accuracy on four independent data sites, reaching 66.9%, 71.3%, 70.8%, and 76.0%, respectively.
- Deep CVAE achieves an accuracy rate of 71.4% using the whole data.
- NYU result is somewhat inferior. Authors state is probably due to the imbalance of the data on this site.

TABLE IV
CLASSIFICATION PERFORMANCE OF DIFFERENT METHODS ON ABIDE DATABASE

Data	Methods	ACC	SEN	SPE	F1
Leuven	PC	57.2±2.9	50.3±4.1	62.9±3.1	51.6±3.6
	SLR	62.8±1.2	**56.6±1.7**	68.0±3.3	**58.0±0.5**
	GC	59.0±3.4	40.7±5.4	75.0±3.4	48.0±5.1
	TE	50.0±1.0	41.4±3.1	57.1±3.1	42.8±1.9
	LiNGAM	59.4±1.0	42.1±4.0	73.7±2.1	48.3±3.0
	DAG-GNN	54.7±2.8	20.7±5.8	**82.9±1.8**	29.0±6.9
	ETLN	47.8±3.8	22.8±6.8	68.6±4.0	28.1±7.5
	CR-VAE	54.1±1.9	27.6±5.3	76.0±2.9	35.0±5.5
	Deep CVAE	**66.9±2.7**	49.7±3.5	81.1±2.9	57.6±3.5
NYU	PC	**66.4±2.1**	**53.4±3.4**	76.2±3.1	**57.7±2.8**
	SLR	63.3±0.8	48.4±3.1	74.5±1.6	53.0±2.0
	GC	52.0±1.2	21.0±3.5	75.2±2.4	27.2±3.6
	TE	58.4±2.6	53.2±3.3	62.3±3.4	52.3±2.9
	LiNGAM	60.0±1.6	45.6±1.1	70.9±2.5	49.5±1.4
	DAG-GNN	58.5±1.2	34.9±2.6	76.2±1.7	41.9±2.3
	ETLN	59.3±1.9	29.9±3.6	**81.5±1.3**	38.6±3.9
	CR-VAE	56.7±1.2	24.3±1.9	81.1±1.5	32.5±2.1
	Deep CVAE	63.9±1.0	43.5±2.2	79.2±0.7	50.9±1.9
UCLA	PC	66.7±1.3	75.1±2.2	57.0±2.9	70.7±1.2
	SLR	60.8±1.2	69.8±2.9	50.4±0.9	65.6±1.7
	GC	58.6±2.3	77.0±3.9	37.4±1.6	66.5±2.3
	TE	48.5±1.9	59.2±2.3	36.1±2.2	55.2±1.8
	LiNGAM	48.3±2.4	54.3±1.9	41.3±3.9	53.0±1.9
	DAG-GNN	52.3±2.8	67.2±4.2	35.2±4.4	60.1±2.6
	ETLN	48.3±2.7	65.3±3.5	28.7±3.7	57.5±2.5
	CR-VAE	56.8±1.6	68.3±1.8	43.5±2.7	62.8±1.3
	Deep CVAE	**71.3±2.0**	**80.8±2.8**	**60.4±2.9**	**75.1±1.9**
UM	PC	66.3±0.5	60.3±1.3	71.7±1.5	62.7±0.6
	SLR	67.3±1.6	**66.5±2.2**	68.1±2.5	**65.6±1.6**
	GC	53.0±2.3	37.0±3.3	67.2±2.5	42.6±3.3
	TE	58.9±0.7	56.8±2.0	60.8±2.7	56.4±0.8
	LiNGAM	58.8±2.2	54.7±3.5	62.3±1.8	55.4±2.8
	DAG-GNN	62.5±2.5	42.9±4.5	79.7±1.8	51.7±4.2
	ETLN	63.4±3.4	48.5±5.0	76.6±3.2	55.4±4.8
	CR-VAE	57.4±1.3	42.4±2.2	70.6±1.0	48.2±2.0
	Deep CVAE	**70.8±1.8**	58.8±3.1	**81.3±1.0**	65.3±2.5
USM	PC	69.3±2.1	72.8±2.0	**64.7±4.0**	73.1±1.7
	SLR	60.4±2.6	69.7±3.9	47.9±3.2	66.9±2.6
	GC	62.2±2.2	**92.6±2.1**	21.0±4.4	73.9±1.4
	TE	59.8±1.5	65.9±2.8	51.6±3.7	65.3±1.6
	LiNGAM	68.7±0.8	77.6±1.1	56.7±1.9	74.0±0.6
	DAG-GNN	58.6±1.9	89.3±2.5	17.2±2.4	71.2±1.5
	ETLN	61.0±1.5	85.9±1.7	27.4±1.7	71.7±1.1
	CR-VAE	53.1±1.0	82.4±0.7	13.5±2.3	66.9+0.6
	Deep CVAE	**76.0±1.9**	85.5±1.8	63.3±3.4	**80.4±1.5**
Whole data	PC	66.4±0.7	63.3±1.2	69.3±1.1	64.6±0.8
	SLR	65.4±0.7	62.2±1.6	68.5±0.9	63.5±1.0
	GC	53.7±1.3	51.2±1.5	56.1±1.7	51.8±1.3
	TE	57.3±1.1	55.8±2.0	58.6±0.6	55.8±1.5
	LiNGAM	58.9±1.6	54.8±1.7	62.8±1.7	56.3±1.7
	DAG-GNN	58.3±1.0	52.2±1.8	63.9±1.2	54.7±1.4
	ETLN	61.0±0.6	58.3±0.9	63.7±0.7	59.1+0.7
	CR-VAE	54.4±1.2	47.0±1.0	61.3±2.2	50.0+1.0
	Deep CVAE	**71.4±0.6**	**66.7±1.0**	**75.8±0.4**	**69.3±0.7**

Ablation Tests

•Designed several degraded networks to confirm whether each component helps to improve the classification performance.

•

1.“CVAE” - use only one hidden layer in each encoder and decoder

2.“deep CVAE_1” - uses four hidden layer in each encoder and decoder

3.“deep CVAE_2” - switch the identity mapping and MLP within each encoder/decoder

•Accuracy of the ADNI and ABIDE datasets does not improve after increasing layers -> not meaningful to increase layers

•Switching identity mapping and MLP -> most detrimental to the accuracy.

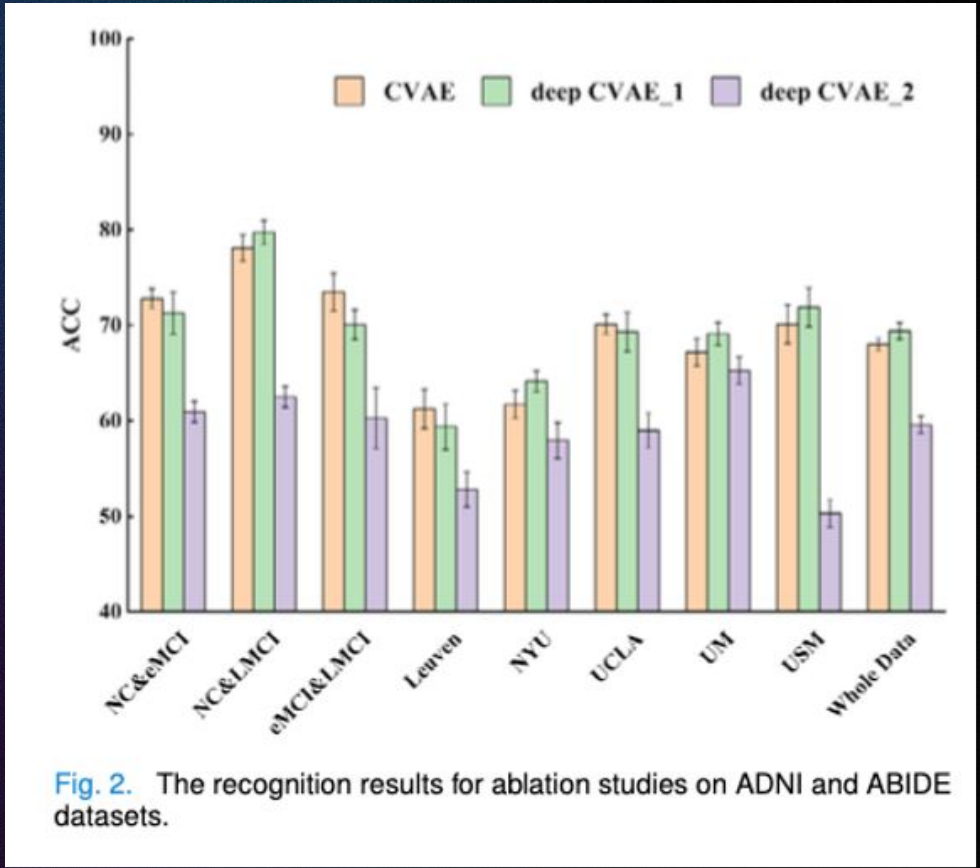
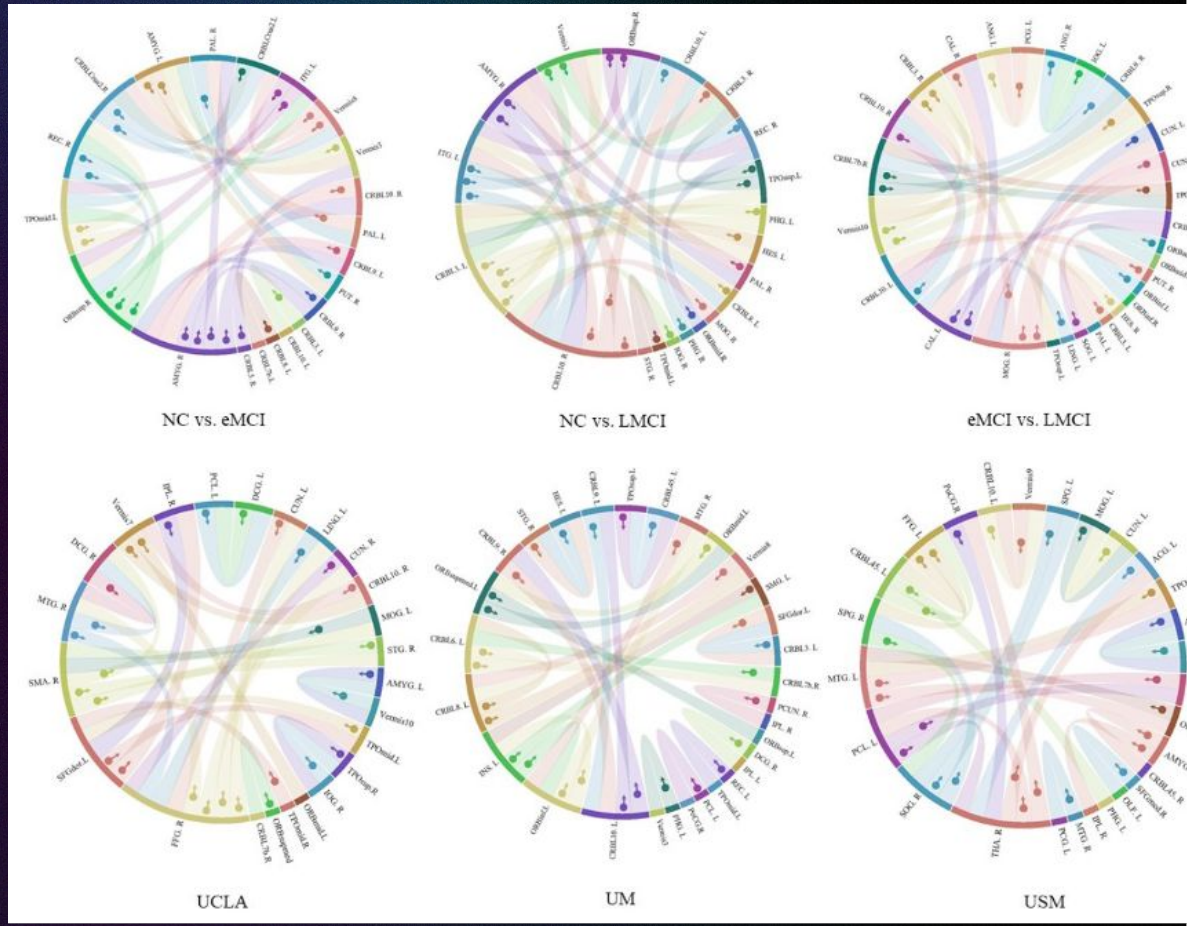


Fig. 2. The recognition results for ablation studies on ADNI and ABIDE datasets.

Most discriminative causal patterns

- Ball - the causal relationship
- Arc width - connection strength
- AD biomarkers
  - CRBL3
  - CRBL10
  - TPOmid.L
- ASD biomarkers
  - MTG.R
  - IPL.R
  - PCL.L



Hierarchical relationships

- Plotted the most discriminative features extracted by the middle layer of the encoder
- Closer to red -> higher group difference
- ADNI
  - Abnormal regions for NC vs. eMCI and NC vs. LMCI
    - frontal lobe
    - temporal lobe
    - parietal lobe
  - Abnormal regions for eMCI vs. LMCI
    - frontal lobe
    - occipital lobe
- ABIDE
  - Frontal, parietal, and temporal lobes
  - Insula

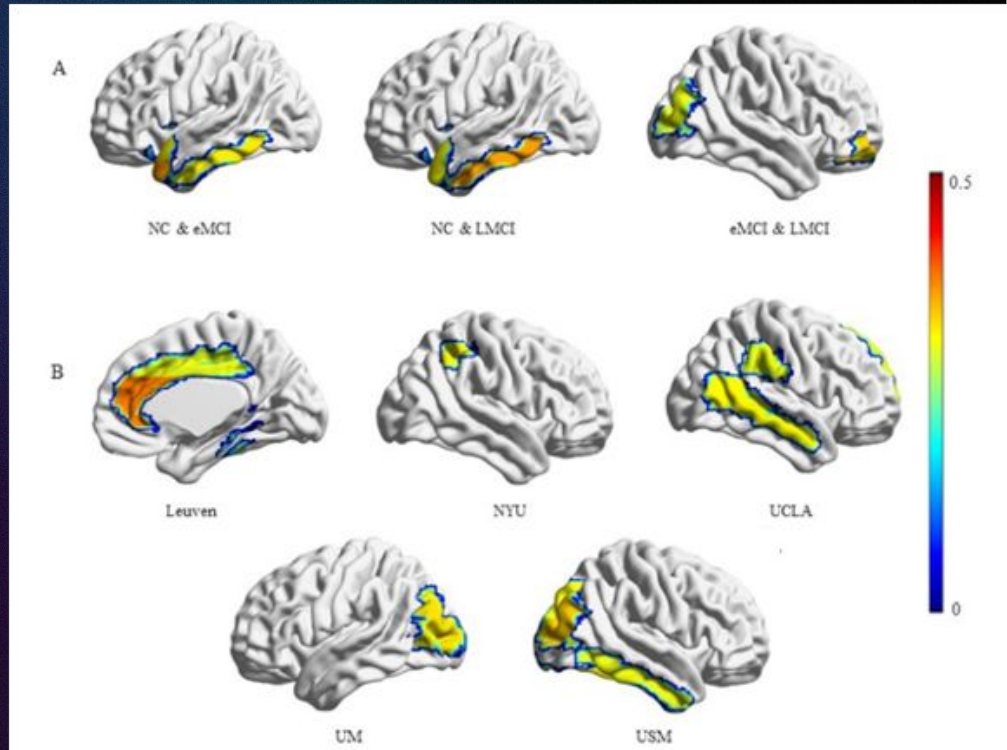


Fig. 4. The hierarchical relationship learned by middle layers of deep CVAE (inference model). (A) for ADNI tasks and (B) for ABIDE tasks.

Reconstruction Error

- Looked at reconstruction errors among groups. Red means higher reconstruction error (abnormality) and blue is lower error.
- ASD regions with highest reconstruction errors
  - frontal lobe
  - parietal lobe
  - insula
- Dementia regions with highest reconstruction errors
  - frontal lobe
  - temporal lobe
  - occipital lobe

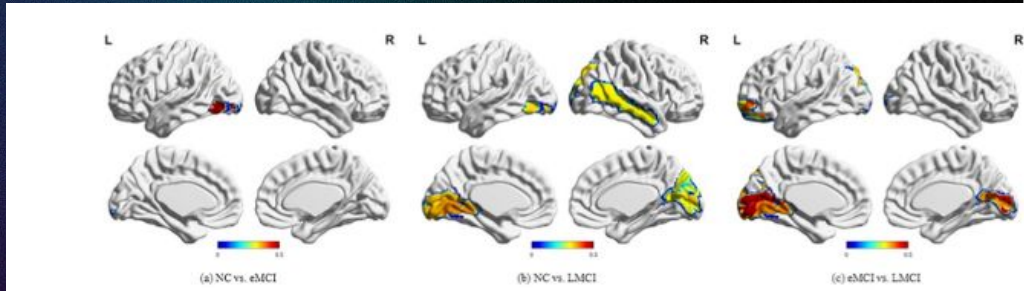


Fig. 5. Dementia-related brain maps. (a) NC vs. eMCI. (b) NC vs. LMCI. (c) eMCI vs. LMCI.

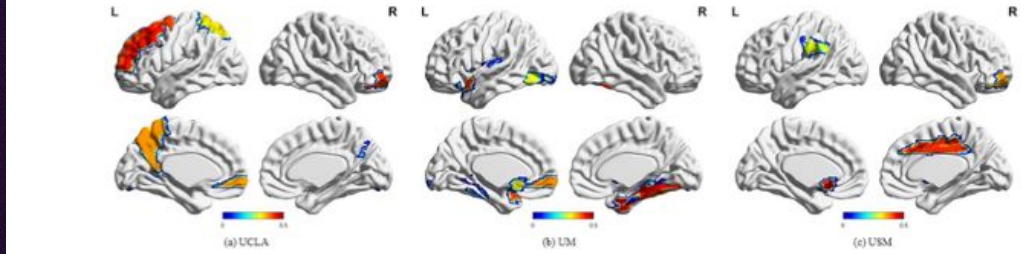


Fig. 6. ASD-related brain maps. (a) Leuven site. (b) USM site. (c) UCLA site.

Conclusion

- They designed (deep CVAE) to estimate causal effects between brain regions
- They achieved good performance on two public databases ADNI and ABIDE

Future Work:

- Could see if results are strengthened at combined timescales
- Could try a denoising VAE to improve the result
- Could explore the learned causal manifold and intervention to possibly change outcomes

References

- https://www.youtube.com/watch?v=eKkESdyMG9w&t=342s
- https://deeplearning.cs.cmu.edu/S20/document/recitation/recitation12.pdf
- https://www.cs.cmu.edu/~bhiksha/courses/deeplearning/Spring.2018/www/slides/lec16.vae.pdf
- https://arxiv.org/pdf/2101.00734
- http://sji.soc.uconn.edu/teaching/deep-learning/index.html