Seminar Report (WiSe 2019)
Visualization and Medical Image Analysis

Using Generative-Discriminative Learning in Neuroimaging for Interpretable Predictions

Shreya Kapoor
Bonn-Aachen International Center for Information Technology

Abstract

Machine Learning has been used to solve a variety of tasks when it comes to Neuroimaging: it can be used to predict the onset of a disease or to classify subjects into healthy and control groups. Even though the models can make predictions with high accuracy, they do not highlight any mechanisms on the basis of which predictions are made. The focus of this seminar report is to explore the state of the art methodology designed to gain mechanistic interpretability of predictions from Machine learning models trained for discriminating from MRI data. A combination of Generative-Discriminative approaches offer ways to incorporate Neurologically important features for patient stratification: the feature extraction is done with the help of a Generative model and classification is done using a Discriminative model. This approach can be implemented in a variety of ways by either using classical Machine Learning or Deep Learning. One particular approach uses Generative embeddings that encode prior knowledge from Neurobiology and then a classifier such as Support Vector Machines (SVMs). Another approach uses Deep Belief Networks to learn latent representations of the imaging data and then feeds it to a classification layer (Softmax in the case of ***Figure 3****). Both methods achieve better performance as compared to traditional classification approaches which are not preceded by a Generative Modeling step and elucidate how models can learn physiologically relevant representations that can highlight the mechanism of a disease.*

1. Introduction

In recent years, MRI (Magnetic Resonance Imaging) has become the preferred modality for Neuroimaging studies due to its safety and ability to create sophisticated 3-dimensional pictures of the brain. In medical practice, it is often used as a tool to visually detect differences between diseased and healthy subjects while testing for Neurological disorders. However, the limitations of the human eye and perceptual vision limit the precision with which medical doctors can predict the presence and cause of a Neurological disorder in a subject. Machine Learning is a good starting point to deal with such classification problems because of its ability to successfully deal with images (example: Segmentation of medical images using 3D U-Nets [ÇAL*16]) and extract features from high dimensional data. However, classification based on Machine Learning is often an artifact of the numerical properties of the data and usually does not incorporate any prior knowledge about the biological process or consideration about system dynamics. In some cases, traditional classification approaches (such as SVMs) were able to achieve good diagnostic accuracy but their results do not offer any insight into the mechanism of the disease under consideration [KSC*08]. To address these challenges it is required to look into the facets of the classification problem being studied.

The structure and task of the discriminative models differs based on the imaging modality used. The two most widely used modalities in Neuroimaging are structural MRI (sMRI) and functional MRI (fMRI, section 2.1). The temporal component and comparatively lower resolution of fMRI is what distinguishes it from sMRI. sMRI scans are used to classify subjects depending upon their anatomical differences while fMRI scans are used to discriminate on with help of variation in the activity of different subjects performing the same task. It is important to get an overview of the different analysis techniques to understand their limitations and scope with respect to interpretable predictions.

Analysis of fMRI images to detect abnormalities in brain function is seeing a paradigm shift from univariate to multivariate methods. The univariate methods (such as Statistical Parametric Mapping) study the statistical relationships between experimental variables and intensity values from each voxel individually [FHW*94]. However, they are unable to detect latent features from the data such as inter-regional connection strengths and globally distributed patterns of activity. Furthermore, this approach ignores the fact that activities of individual voxels are not independent of each other [KMD*09]. To address the limitations of the univariate methods, multivariate ones incorporate information from an ensemble of voxels and in this way encode connectivity information. Examples of multivariate classifiers that can be trained to classify from fMRI scans are nearest-neighbors classifier, Fisher's linear discriminant, Gaussian Naïve Bayes, and linear and nonlinear (radial-basis-function kernel) SVMs. sMRI scans are usually treated as 3-dimensional images and anatomical basis of a disease is analysed

by taking into account differences in the specific Regions of Interest (ROIs). Patients and healthy controls are assumed to have different levels of activity in brain regions considered to have different functionality. A set of promising algorithms that could extract features from specific regions of interest are 3D CNNs [KAB*18]. Another type of Neural Network model, the Deep Belief Networks (DBNs) [Hin] can be used to extract statistically relevant information from an ensemble of voxels represented in an sMRI scan. [PHS*14]

The remainder of this report is structured in the following manner. First, the background section gives a brief introduction to the techniques used to implement the pipelines illustrated in the methods section. This section gives an insight into the different pipelines used for patient stratification based on fMRI and sMRI scans respectively. Both pipelines incorporate a combination of Generative-Discriminative learning but use different ways. Third, the results from the two different approaches and their prospects. Finally, the Discussion section illustrates individual opinion and analysis of two different approaches. The concluding remarks highlight the extent to which the challenges stated have been overcome.

2. Background

2.1. Magnetic Resonance Imaging

MRI is a non-invasive imaging technology that produces 3-dimensional detailed anatomical images [MMGP06]. This technique is based on the Nuclear Magnetic Resonance Imaging principle. It uses the difference between the magnetic properties of different tissue types presented with an external magnetic field to generate images.

There are two types of MRI scans: Structural (sMRI) and functional (fMRI). Structural MRI consists of 3-dimensional high resolution anatomical images of the brain and are often used to study the locations of brain defects in patients. On the other hand, functional MRI scans are 4-dimensional tensors with time as the fourth dimension. They are often used for task-based Neuroscience experiments designed to study differences of activity between groups of subjects performing the task or tracking activity patterns through the course of time.

2.2. Unsupervised Feature Learning with Generative Modelling

Generative modelling usually refers to inferring the probability distribution underlying the given data and generating new samples from it. It enables model inference by capturing the joint probability distribution P(X,Y) of the data X and the labels Y (Supervised learning) or only P(X) if the labels are absent (Unsupervised learning). Some of the commonly used generative models for unsupervised learning tasks are Gaussian Mixture models and Hidden Markov Models. Another set of models, such as DCM (section 2.2.1) and Restricted Boltzmann Machines (section 2.2.2) are probability based and relevant to unsupervised feature learning tasks. They learn useful transformations of the original data which can serve as preprocessing steps or help low dimensional exploration.

Unsupervised learning is important for MRI based studies because the raw data in the voxel space is high dimensional (about 1 million) and the individual voxels themselves are much less significant than a cluster of voxels for predictive tasks. In the following subsections, two different approaches to achieve unsupervised feature learning have been highlighted. First one uses Dynamic Causal Modelling to include prior knowledge from Neurobiology and the second one uses Boltzmann state function to extract features without any prior knowledge.

2.2.1. Dynamic Causal Models (DCMs)

DCM is a general framework for inferring processes and mechanisms at the neuronal level from measurements of brain activity [FHP03]. These are often used with modalities such as fMRI that have a specific time component. In this framework, each brain region has an input, state and output. The activity of neural populations in each pre-specified brain region is represented by a single state variable x and is perturbed by (known) experimental stimuli u. Suppose a function f models the change of state of the neural state vector x (state vector, reflecting the brain state by the ensemble activity at a given time) dependent on internal parameters and stimulus such that, $f(0,0) = 0$. The dynamics of the neuronal system are modelled with the help of a non-linear function f such that it can be approximated with the help of a Taylor series truncated until second order differentials [SKH*08].

$$f(x,u) = \frac{\delta x}{\delta t} \approx f(0,0) + \frac{\partial f}{\partial x}x + \frac{\partial f}{\partial u}u + \frac{\partial^2 f}{\partial x \partial u}xu \tag{1}$$

$$f(x,u) = \frac{\delta x}{\delta t} = (A + \sum_{i=1}^{M} u_i B^{(i)})x + Cu \tag{2}$$

The bilinear term describes the interactions between neuronal states x and inputs u. Given m known (stimulus) inputs, one can parameterize the equation 1 with $A = \frac{\partial f}{\partial x}|_{u=0}$, $B(i) = \frac{\partial^2 f}{\partial x \partial u_{(i)}}$, and$C = \frac{\partial f}{\partial u}|_{x=0}$ to obtain the form in equation 2. The parameters are $\theta_n = (A, B^1, B^2, ..., B^j, C)$ rate constants with the units of frequency. In equation 2, the matrix A represents the fixed (context-independent or endogenous) strength of connections between the modelled regions, and the matrices $B^{(i)}$ represent the changes of fixed connections induced by the i_{th} input u_i which adds on to changes of the dynamic system. Finally, the C matrix represents the influence of direct (exogenous) inputs to the system (e.g. sensory stimuli). In this manner, the DCM helps to explain the dynamics of the brain while taking into account internal parameters and changes induced due to external stimuli.

The inversion of a DCM [BSL*11] gives the distribution of $p(\theta|X,m)$ where m = model (subject-specific with inter-regional connection strength parameters), and X are the measurements from an individual subject. From this distribution the parameters with the Maximum a posterior probability (MAP) estimate gives an approximation about the subject-specific inter-regional connection strengths.

2.2.2. Restricted Boltzmann Machines and Deep Belief Networks

An RBM(Restricted Boltzmann Machine) is a generative neural network consisting of only two layers; a visible and a hidden layer.

The two layers have symmetric connections and none within them. The hidden units can be viewed as non-linear feature vectors detectors aiming to model the dependencies of the inputs. [FI]

When RBMs are used for extracting features from MRI data, the samples in the voxel space are fed into the visible units and are mapped to the reduced feature space defined by the hidden layer. It has been shown that RBMs can learn transformations that identify networks and their temporal activations from fMRI data [HCS*14].

In general, the weights learned by the network determine the transformation from the visible layer to the hidden layer. The joint distribution learned by the model is given by $p(v,h) = \frac{1}{Z}e - E(v,h)$ where v is the visible layer vector(input vector) and h is the hidden layer vector, Z is the partition function and E is the energy of the network given by

$$E(v,h) = -\sum_{i=1}^{N}\sum_{j=1}^{M} w_{ij}h_i v_j - \sum_{i=1}^{N} c_i h_i - \sum_{j=1}^{M} b_j v_j \tag{3}$$

W_{ij} is the weight between the hidden unit i and the visible unit j. The b's are the bias terms of the visible units and c's are the bias terms of the hidden units. M, N represent the number of visible and hidden layer units respectively.

The minimization of the energy leads to maximisation of the probability and the maximal probability distribution is what gives us the best parameters describing the data The training of an RBM consists of two steps:

- **Gibbs Sampling**: a hidden layer sample can be obtained based on $p(h|v)$ and visible layer sample $p(v|h)$. This is known as block Gibbs sampling. Considering the sigmoid as the activation function the probabilities are given by

$$p(h_i = 1|v) = \frac{1}{1+e^{-c_i+\sum_j^M W_{ij}v_j}} \tag{4}$$

$$p(v_j = 1|h) = \frac{1}{1+e^{-b_j+\sum_i^N W_{ij}h_i}} \tag{5}$$

  Using these probabilities the log likelihood of the configuration can be easily determined and maximised.
- **Contrastive Divergence** for approximating the log likelihood gradient:

$$CD_k = -\sum_h p(h|v^{(0)})\frac{\partial E(v^{(0)},h)}{\partial\theta} + \sum_h p(v|h^k)\frac{\partial E(v^{(k)},h)}{\partial\theta} \tag{6}$$

  The above formulation in equation 6 iteratively minimizes the energy between the visible and the hidden layer. The Gibbs chain is carried out with the help of a number of pre-specified steps k and initialized with $v^{(0)}$. The subsequent $h^{(t)}$ is sampled from $p(h|v^{(t)})$ and similarly $v^{(t+1)}$ from $p(v|h^{(t)})$ where $t = 1..k$. With the help of the gradient calculations the weight updates can be carried out an optimised. The weight matrix learned after k iterations becomes the transformation from the original feature space to the reduced feature space.
  An extension of RBM is a Deep Belief Network (DBN). It is a neural network which consists of stacked RBM layers. This stacking enables the algorithm to learn more complex and non-linear transformations. Usually addition of layers increases the performance and the non-linearity in the network. DBNs are trained just like any other neural network with the help of back-propagation [Hin].

2.3. Combination of Generative and Discriminative approaches

It is well known that discriminative models aim to learn a mapping from the input feature space to the labels and do not essentially focus on providing human interpretable decision rules. Even the performance metrics of traditional classifiers do not provide an insight into how the model reached its conclusion [Dav19]. The discriminative models could also learn unimportant features that are just numerically more represented and might not be mechanistically important. That is why a combination of Generative-Discriminative learning is needed in order to stratify patients in a clinical setting. In this type of learning, the Generative model serves as a sort of pre-processing for the classification step. The induced feature space is lower dimensional and more informative and this very property enables interpretability of classification results. Such models, like Conditional Variational Autoencoders (cVAEs) [YRYR19] have been well implemented for handwritten images but their applicability to the medical domain remains in its nascent stages [Dav19].

2.4. Embedding high dimensional feature space for exploratory analysis

2.4.1. Generative kernels

In this section, we mention Generative kernels to introduce ideas important to understand the methodology in section 3.1. Let us suppose with given sample data X, the Generative model learns the probability distribution $P(X)$ having parameters θ which belong to a parametric family M_Θ (the set of all possible values that the parameters can take). The Generative kernels implement functions that define a similarity metric for observed examples. [BSL*11]. The main intuition is to derive the kernel functions from a generative probability model [JLH]. The Generative embeddings make use of this kernels for classification. The kernel constructs a model based feature space in which samples are represented by their statistical properties. In the above mentioned case, this helps the classifier to learn from the Neurobiologically important features (the θ parameters) learned with the help of the DCM.

2.4.2. Constraint Satisfaction Problem with Divide and Conquer

A non-linear embedding approach can be implemented with the help of a Constraint Satisfaction Problem (CSP). This type of approach has been used in section 3.2 in order to visualise and interpret the feature space learned by a Generative model. In general, a CSP is one in which the solution is required to meet previously specified constraints (such as preservation of neighborhood relations or distribution of the raw data). A DC (Divide and Conquer) algorithm can be used to solve CSPs [GE08]. The DC strategy breaks down the bigger problem (Constraints on the embedding of the whole dataset) into smaller sub-problems, first solve the smaller sub-problems and use it to get the solution to the bigger problem.

For the methodology proposed in 3.2 consider the d-dimensional

embeddings.png

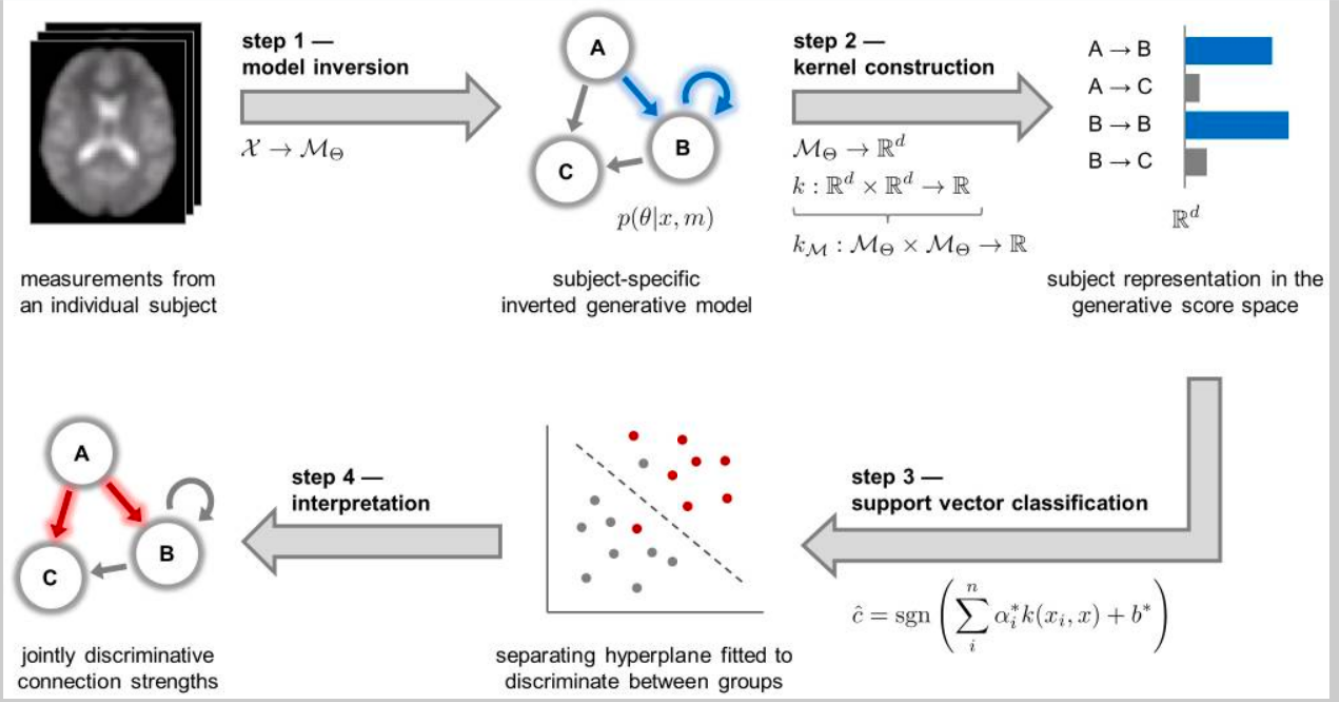


Figure 1: ***The pipeline of Generative Embedding for Model-Based Classification of fMRI data****. The detailed explanation of the pipeline is given in section 3.1. At first, the BOLD (blood oxygen level dependent) activity of each subject is represented by an fMRI timeseries and then the DCM (section 2.2.1) of the thalamo-temporal cortex is inverted to get the values of the parameters for each subject. For this particular subject, the connections between the regions and A and B and the self connections of region B are more illustrative than others. The second step is the kernel construction explained in section 2.4.1. The linear kernel is used to achieve linear separability in the higher dimensional space and the features correspond to the coupling strengths between difference regions. The feature weights tell us which connections are more important than others and in the case of this particular dataset the connections between A and B, A and C are more important than other connections. Note: A,B,C in this figure correspond to brain regions and are different from A,B,C in section 2.2.1.* ***Source: [BSL*11]***

hidden node space. The embedding method wants to generate a good mapping from d-dimensional to a desired low dimensional space. For this, each point in the solution map (i.e. in the 2D map) is taken to be a variable and has some associated constraints. Each point gets a copy for each constraint (each variable has one constraint, one point gets copies along n constraints/dimensions where n = number of points). The divide strategy moves the points in 2D which are closer together (for each point, the algorithm determines which of the replicas satisfy the constraints and then brings these closer) while the conquer averages location of all replicas (nearest neighbors).

3. Methods

Classification of medical images needs to be treated as a special case of image classification since they are used to guide clinical decisions. Traditional discriminative approaches give predictions on the basis of a learned mapping from the input space to output classes and do not always offer mechanistic interpretability. A combination of Generative and Discriminative methods (section 2.3) is required for building robust classifiers which make decisions on the basis of latent characteristics of the data. Here, the Generative model serves as a type of preprocessing or dimensionality reduction step that generates a comparatively Neurobiologically intuitive feature space on the basis of which the classifier makes the decision [Dav19].

3.1. Classification of fMRI data using Generative embeddings

In the paper titled 'Generative Embedding for Model-Based Classification of fMRI Data' [BSL*11] Generative embeddings were used to learn from an fMRI experiment of subjects performing speech processsing tasks. The study incorporated data from $n = 37$ subjects, 26 healthy and 11 diagnosed with moderate aphasia. In the first step of the pipeline given in **Figure** 1, the generative model used is a DCM (section 2.2.1). It gives the change of the state vector of the Neuronal system (with respect to time) based on strength of inter-regional connections. It is based on the assumption that the brain is a dynamic system whose activity is a function of both external stimulus and internal connections. The parameters of the DCM (equation 1) give the coefficients between brain dynamics and internal states as well as brain dynamics and output stimulus. These very coefficients remain unknown due to high network complexity of the brain and need to be derived experimentally. In this case, a

Region		MNI coordinates
L.MGB	left medial geniculate body	−23 mm, −23 mm, −1 mm
L.HG	left Heschl's gyrus (A1)	−47 mm, −26 mm, 7 mm
L.PT	left planum temporale	−64 mm, −23 mm, 8 mm
R.MGB	right medial geniculate body	22 mm, −21 mm, −1 mm
R.HG	right Heschl's gyrus (A1)	48 mm, −24 mm, 6 mm
R.PT	right planum temporale	65 mm, −22 mm, 3 mm

Speech processing can be modelled using a dynamic causal model (DCM) with 6 regions. The table lists the central coordinates of these regions in MNI152 space. These coordinates define the centre of the rough anatomical masks (16 mm×16 mm×16 mm) that guided the specification of the exact location and extent of the regions of interest underlying model inversion (see Section 'Implementation of generative embedding'). For an illustration of these masks, see Figure S1 in the Supplementary Material.
doi:10.1371/journal.pcbi.1002079.t001

Figure 2: ***Regions of interest(ROIs)*** *in the Thalamo-temporal cortex which are essential to model speech processing tasks. The network representation used for testing differences between aphasic patients and healthy controls is shown in figure 6* ***Source:*** *[BSL*11]*

DCM for speech processing tasks balancing accuracy and model complexity (section 6) was inverted in order to obtain strength of inter-regional connections specific to one subject. This inversion step mapped the data X to a multivariate probability distribution $p(\theta|x,m)$ in a parametric family M_Θ and from this distribution the Maximum a posterior probability (MAP) estimate was taken . The model inversion followed by the mapping $M_\Theta \rightarrow R^d$ is the **Generative Embedding**.

For the achieving the goals stated in the step above, DCM of the thalomo-temporal cortex with 6 Regions of Interest(ROIs) were used to measure differences between aphasic and non-aphasic subjects in speech processing tasks. It is well known that such ROIs are defined in order to gain exploratory direction, statistical control or functional specification in fMRI analysis [Pol07]. The second step involved generating a score space(with the help of a generative kernel) from a MAP estimateand then using simple linear kernel [PC13] to compute pairwise dot products between subjects. It can be seen from **Figure** 1 that in step 2, first the best parameters for each subject were mapped to a d-dimensional space (defined by the DCM where d is the number of connections in the model). The mapping used is a function $f : M_\Theta \rightarrow R^d$ that extracts μ_{MAP} from a subset of MAP estimates in the posterior distribution $P(\theta|X,m)$. Also, the d-dimensional space (called the generative score space) encoded inter-neuronal synaptic connections and not activity in different brain regions. Once the generative score space had been created then a linear kernel was used to compute dot products of the statistical representations of each subject (the parameters of their inverted DCMs). The linear kernel function can be represented using the equation $R^d x R^d \rightarrow R$.

To summarise the second step: after representing one subject in the score space a linear Kernel is used to compute scalar dot products between parameters (inter-regional connection strengths) of each subject. The whole step can be viewed as generating a probability kernel $K_M : M_\Theta x M_\Theta \rightarrow R$ that gives a measure about the similarity of parameters of two subjects when using a particular DCM. In the third step, the dot products from the Generative kernel were used to solve the optimisation problem of an SVM classifier and find an optimally separating hyperplane [BB98]. Finally, in the fourth step the constructed feature space was investigated to infer about parameters that jointly give the most discriminative information between healthy subjects and controls. The interpretation can be done on the basis of numerical values of the feature weights. A feature with a higher weight was expected to contribute more numerically than lower weighted features. This was used to infer the highly relevant inter-regional connection strengths for classification.

3.2. Classification of sMRI data using Deep Belief Networks

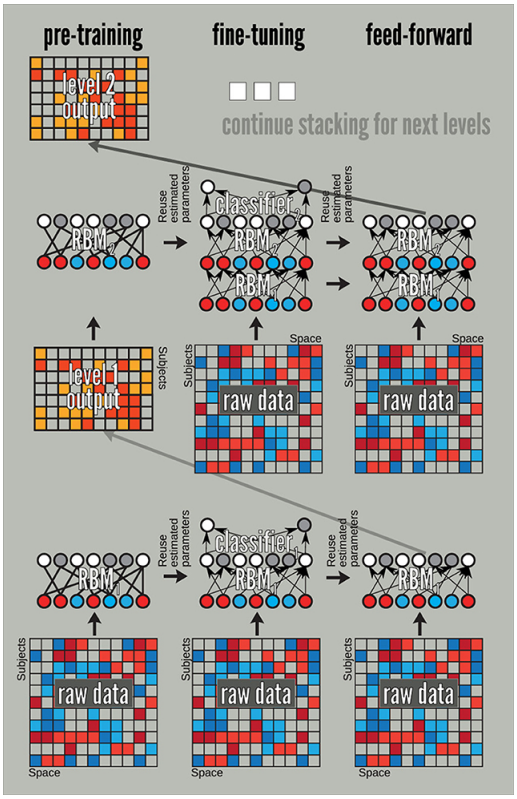


Figure 3: ***Training of a Deep Belief Network*** *with 3 layers. The DBN is treated as a consecutive stacking of RBMs. It can be seen that the first RBM consists of the visible units receive data from the voxel space and a hidden layer, which then serves as the visible unti input for the second RBM and so on. For this particular study a 50-50-100 architecture was used i.e. 50 neurons in first and second hidden layers with 100 neurons in the third hidden layer. The 100 dimensional-input patterns in learned feature space are then fed to a classifier layer (softmax in figure) that makes a yes/no decision.* ***Source:*** *[PHS*14]*

The pipeline illustrated in **Figure** 3 was used for two tasks. First, to model the progression of Huntington disease on the basis of sMRI scans. Second, to analyse the effect the depth of the DBN effect on classification on a Schizophrenia sMRI dataset.

Depth	Pre-training				Fine tuning			
	Input	1	2	3	Input	1	2	3
Dimension	60465	50	50	100	60465	50	50	100
Unit type	Gaussian	Logistic	Logistic	Logistic	–	Logistic	Logistic	Logistic
Dropout probability	0.2	0.5	0.5	0.5	0.7	0.5	0.5	0.75
L_1 Regularization	–	0.1	0.01	0.001	–	0.001	–	–
Learning rate	–	0.01	0.01	0.001	–	0.01	0.1	1e–8

Figure 4: ***Parameter ranges*** *used for pre-training and fine tuning in order to train the different depth DBNs in section 3.2.2.* ***Source:*** *[PHS*14]*

For both the tasks above, the training of the DBN was done in a greedy manner by treating successive layers as RBMs. For combination with a classification approach, RBM training was divided in two steps:pre-training, discriminative fine-tuning.

- **Pre-training:** The input from the previous layer or the raw data (in case of layer 1) was fed to the visible layer of the RBM.
- **Discriminative fine-tuning:** The parameters from the pre-training stage along with the raw data were fed to the network in order to get the hidden-layer values. The hidden layer values were then fed to a classification layer. This step allowed the training of the RBM as a feedforward neural network with back-propagation of error. In this particular case, the soft-max layer had access to only the binary labels: diseased and healthy.

Another important aspect of this implementation that helps to guide interpretation of the classification is to incorporate a non-linear embedding method to control what the model learns in each layer. This was done with the help of a Constraint Satisfaction Framework(section 2.4.2). The CSP is a method to inspect the data samples in the hidden layer feature space. A 2D representation of the feature space in each layer is computed with the help of a divide and conquer strategy such that it preserves neighborhood relations and helps get an intuition about the network learning the correct properties about relationships between the data samples being preserved after transformation. The DBN output was then visualised with the help of the CSP that preserves neighborhood relationships from the higher dimensional feature space (100 dimensional in the case of the depth 3 model) in the 2-dimensional space.

3.2.1. Huntington Disease dataset

This dataset involved analysis of 3500 sMRI scans (2641 patients and 859 controls) from project PREDICT-HD (www.predict-hd.net) [LPP15] where it was tested to guide knowledge discovery about changes in cognitive skills when a patient transitions from healthy to a diseased state. In this case, a three layer DBN was trained in an unsupervised manner for feature learning and the additional softmax layer was used for the classification task. The visible layer of the DBN received input from the voxel space (with a pre-defined rule concerned with regions of interest) and the hidden space is the learnt feature space.

Measure($n=37$)	Searchlight feature selection	PCA-based-dimensionality reduction	Generative-embedding(full-model)
Accuracy	0.730	0.865	0.973
Balanced Accuracy	0.729	0.799	0.981
Significantly above chance	p =.006	p<0.001	p<0.001

Table 1: ***Results of classification with linear Kernel SVM*** *from the Generative embedding approach when compared to PCA-based dimensionality reduction [WEG87] and Searchlight feature reduction [KGB06]. A few of the results from the original paper have been presented to illustrate the difference Generative embedding makes as a pre-processing step to increase the balanced accuracy.* ***Source: Table 2 [BSL*11]***

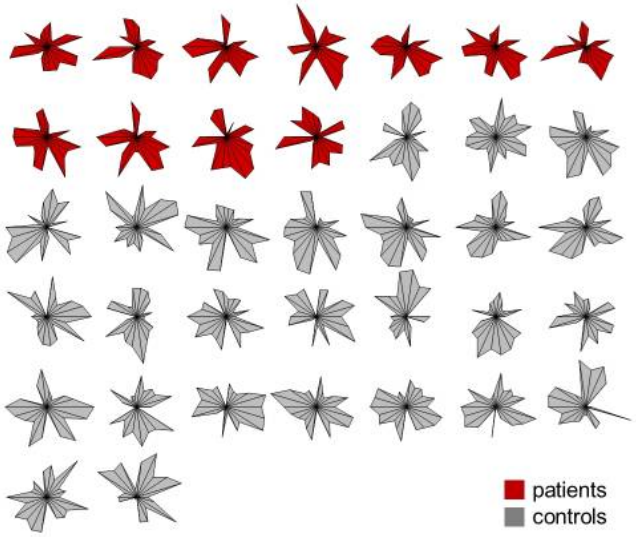
3.2.2. Schizophrenia dataset

The different variants(1, 2 and 3 layer DBNs) were used for classification on a dataset combined from four studies at John's Hopkins University(JHU) with 198 schizophrenia patients(both first episode and chronic patients) and 191 matched healthy controls. To analyse the effect of depth in the DBN, three different model architectures were used. First, an RBM with 50 hidden units in the top layer. Second, a DBN with depth 2 with 50 units in first layer and 50 in the top layer. Third, a DBN of depth 3 with 50,50-100 units in first,second and top layers. The parameter ranges which were used in order to evaluate the depth effect are put together in **Figure 4**. Logistic regression, SVM and kNN classifiers used the DBNs as the pre-processing step and the F-scores were recorded accordingly.

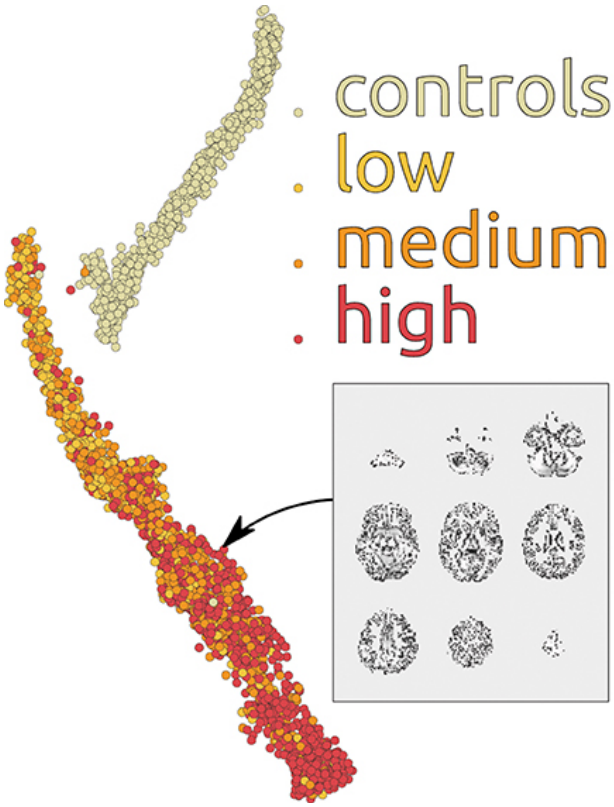
4. Results

4.1. Generative Embedding Approach

Using the DCM for speech processing **Figure 6** the three most important connections that help to distinguish between patients and controls are: R.PT to L.PT, L.HG to L.PT (both forward and backward connections) and R.HG to L.HG (see **Figure 4**). They are highlighted in bold red and have obtained with cross validation. The directed synaptic connections are all terminating in the left hemi-



(a) **Visualisation of different subjects with the help of radial coordinates**. The axis along each radial coordinate is a parameter of the DCM (representing inter-regional connection strengths) of the thalamo-temporal cortex, the MAP estimates are plotted on the respective axis'. It can be seen that there is not an obvious difference between aphasic (red) and healthy controls(grey). This particular observation highlights the fact that presence of aphasia can be attributed to a combination of the factors rather than any particular factor independently. **Source:** [BSL*11]



(b) **Visualisation of subjects in 2D space using a CSP**. Each point on the map is an sMRI volume which is transformed to 100 dimensional feature space by the DBN, the data in the DBN space is then represented by an embedding in 2D with the help of divide and conquer strategy (section 2.4.2). Even though the network did not have any information about the severity of the disease (the discriminative fine tuning had access to only binary labels, see section 3.2), the decomposition by the CSP could generate a direction corresponding to the disease severity and this signifies that the DBN learns important features for classification. **Source:** [PHS*14]

Figure 5: *Results of Generative modelling from the two approaches*

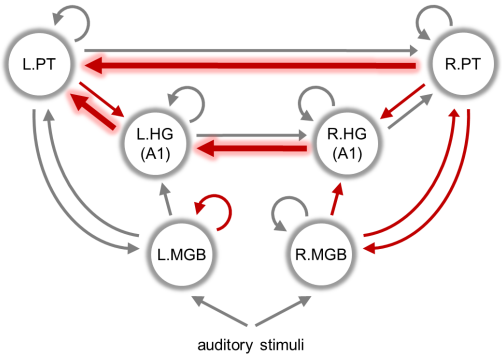


Figure 6: ***Results of DCM interpretation*** *after training the classifier with the pipline in* ***Figure 1*** *the discriminative connections strengths inferred from the dataset have been highlighted in red. The bold red connections are the ones which have significant discriminative power repeatedly after cross-validation. The DCM has 6 regions of interests as mentioned in* ***Figure 2****, 15 inter-regional connections, 6 self connections and two auditory stimulus, it has been selected on the basis of it's ability to balance accuracy and complexity as compared to other models of the thalamo-temporal cortex(Schofield et. al).* ***Source: [BSL*11]***

Depth	Raw	1	2	3
SVM,F-score	0.68 ± 0.01	0.66 ± 0.09	0.62 ± 0.12	0.90 ± 0.14
LR,F-score	0.63 ± 0.09	0.65 ± 0.11	0.61 ± 0.12	0.91 ± 0.14
KNN,F-score	0.61 ± 0.11	0.55 ± 0.15	0.58 ± 0.16	0.90 ± 0.16

Table 2: ***Results of classification using the DBN****. The F-scores from the different layer DBNs have been compared to the raw data. The classification here has been done by using a SVM, LR or KNN after the last hidden layer of the RBM.****Source:*** *[PHS*14]*

sphere and highlight transfer of information from right to left hemisphere. This is in accordance with prior knowledge that language is processed in the brain by the transfer of information from the right to the left hemisphere [SMP*07]. The mechanistic interpretability offered with this type of modelling is the fact that these discriminatory connections illustrate that language is processed differently between aphasic and non aphasic patients and it is the inter-regional forward connections from the right to the left hemisphere which are disrupted for those suffering from aphasia.

With the help of **Figure 5(a)** it can be seen that the model induced generative score space gives an intuitive visualisation and correspondence with the DCM. The 'connectional fingerprints' [PSK02] of fMRI scans illustrate that the difference between patients and controls might not be intuitive at the first glance but probably a combination of parameters could help visually see the differences. Further, the Generative embedding enabled linear SVM classifier to get a balanced accuracy of 98% with a p-value<0.01 which signifies that the results of this model are not random and hold statistical significance. Also, in **Table 4.1** it can be seen using

the Generative embeddings as a pre-processing step has the highest balanced accuracy which is a better performance metric than plain accuracy because the dataset used was imbalanced, i.e. 26 controls and 11 diseased.

4.2. DBN Approach

4.2.1. Huntington Dataset

As exhibited by **Figure 5(b)**, the DBN extracts high level features from the raw sMRI data which are visualised with the help of a non-linear embedding approach. It can be seen that the healthy controls and diseased subjects are well separated in the 2D embedding. Even though the discriminative fine tuning phase did not have access to the disease severity (color-coded by authors on the map) the embedding from the DBN space still output directions corresponding from low to high disease severity. Since the embedding preserves neighborhood relationships from the hidden node space of the DBN (100 dimensional top layer of the 3-layered network) it can extrapolated that the 100 hidden node space is a statistically relevant representation of the voxel space. From the 2D embedding the subjects can also be grouped due to the spectral decomposition (eigenvector obtained by solving the CSP), the 'extent' of difference between any two subjects can be determined via their distance on the 2D map. In a very ideal scenario, the sMRI scans can then be used to detect spatial locations or ROIs (using techniques such as segmentation) which are important for the distance of samples in the embedding.

4.2.2. Schizophrenia Dataset

In all three cases represented by **Figure** 4.1 the DBN based classification showed an increase in the F-scores with the increase in the number of layers. The increasing F-scores illustrate that the DBN possibly learns useful transformations of the original spatial MRI data. With the parameter settings in **Figure 4** it can be seen that Deep Learning can do automatic feature learning from even a large number of features without encoding any prior knowledge about the problem at hand. The statistical formulation of the Boltzmann distribution is what helps the network to detect high non-linearity in the data.

5. Discussion

The two methodologies (section 3) mentioned above combine Generative-Discriminative models in different ways; first(section 3.1) using prior knowledge from Neurobiology and second(section 3.2) without.

The first one gives good results in terms of balanced accuracy and Neurobiological interpretability but the experiment was performed only for 37 patients, a small dataset, the applicability of this particular pipeline remains questionable. The most important reason for the accuracy of this method remains that prior knowledge from Neurobiology encoded in DCMs helps to see a collection of voxels performing activity and not one in isolation. Also, the parameter learning remains unsupervised and it lies on the assumptions that inter-regional connection strengths are more important for discrimination than isolated regions. The last interpretation step in the pipeline 1 helps to see which connections are important and characterise the patients of the disease.

The second methodology uses a Deep Learning approach to extract features from the raw data. This holds lots of promises when the aim is to explore mechanisms of diseases about which no prior knowledge is present. It also removes the subjectivity from the model design (e.g. what type of information a particular research group might want to encode). Also, adding more layers could also help explore higher level complex features in the data which might not have an exact correspondent in Neurobiology but can certainly be spatially visualised with embeddings and weight matrices.

6. Conclusions

The two different methods offer good visualisations and help to see the Neurobiologically important information from different aspects. In summary, the applicability of both of these methods needs to be explored for different types of Neurological disorders affecting different parts of the brain. It is important to note that even though MRI images were used for the two different tasks mentioned in the methodology, the addition of time component makes fMRI feature learning different from sMRI feature learning.

When it comes to comparing the two methods in terms of Neurobiologically interpretability, the Generative embeddings are useful in cases when we already have some prior knowledge about the Neurological disorder; for example which brain area is affected and which connections to model. On the other hand, the DBN approach is helpful in cases where no prior knowledge about the disorder is known and latent features need to be learned in an unsupervised manner. We can say that the Deep learning approach is 'fully automatic feature learning' while the Generative embeddings are 'semi-automatic feature learning'.

With results from Generative-Discriminative learning, one can conclude that it certainly helps to encode information about the brain as a system, an ensemble of voxels. It holds promises for the field of Neuroimaging because of increased classification performance and ability to enable statistical inference as compared to traditional classifiers. From the results presented above, it can be seen that the statistical inference from such models along with problem-specific model design can certainly achieve mechanistic interpretability about a disease.

References

[BB98] BURGES C., BURGES C. J.: A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery 2* (January 1998), 121–167. URL: https://www.microsoft.com/en-us/research/publication/a-tutorial-on-support-vector-machines-for-pattern-recogn 5

[BSL*11] BRODERSEN K. H., SCHOFIELD T. M., LEFF A. P., ONG C. S., LOMAKINA E. I., BUHMANN J. M., STEPHAN K. E.: Generative embedding for model-based classification of fmri data. *PLOS Computational Biology 7*, 6 (06 2011), 1–19. URL: https://doi.org/10.1371/journal.pcbi.1002079, doi:10.1371/journal.pcbi.1002079. 2, 3, 4, 5, 6, 7

[ÇAL*16] ÇIÇEK Ö., ABDULKADIR A., LIENKAMP S. S., BROX T.,

RONNEBERGER O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention* (2016), Springer, pp. 424–432. 1

[Dav19] DAVATZIKOS C.: Machine learning in neuroimaging: Progress and challenges. *NeuroImage 197* (2019), 652 – 656. URL: http://www.sciencedirect.com/science/article/pii/S1053811918319621, doi:https://doi.org/10.1016/j.neuroimage.2018.10.003. 3, 4

[FHP03] FRISTON K. J., HARRISON L., PENNY W.: Dynamic causal modelling. *Neuroimage 19*, 4 (2003), 1273–1302. 2

[FHW*94] FRISTON K. J., HOLMES A. P., WORSLEY K. J., POLINE J.-P., FRITH C. D., FRACKOWIAK R. S.: Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping 2*, 4 (1994), 189–210. 1

[FI] FISCHER A., IGEL C.: An introduction to restricted boltzmann machines. 3

[GE08] GRAVEL S., ELSER V.: Divide and concur: A general approach to constraint satisfaction. *Physical Review E 78*, 3 (2008), 036706. 3

[HCS*14] HJELM R. D., CALHOUN V. D., SALAKHUTDINOV R., ALLEN E. A., ADALI T., PLIS S. M.: Restricted boltzmann machines for neuroimaging: An application in identifying intrinsic networks. *NeuroImage 96* (2014), 245 – 260. URL: http://www.sciencedirect.com/science/article/pii/S1053811914002080, doi:https://doi.org/10.1016/j.neuroimage.2014.03.048. 3

[Hin] HINTON G. E.: Deep belief nets. 2, 3

[JLH] JAAKKOLA T. S., LABORATORIO M. A. I., HAUSSLER D.: Exploiting generative models discriminative classifiers. 3

[KAB*18] KHVOSTIKOV A., ADERGHAL K., BENOIS-PINEAU J., KRYLOV A. S., CATHELINE G.: 3d cnn-based classification using smri and MD-DTI images for alzheimer disease studies. *CoRR abs/1801.05968* (2018). URL: http://arxiv.org/abs/1801.05968, arXiv:1801.05968. 2

[KGB06] KRIEGESKORTE N., GOEBEL R., BANDETTINI P.: Information-based functional brain mapping. *Proceedings of the National Academy of Sciences 103*, 10 (2006), 3863–3868. 6

[KMD*09] KOUTSOULERIS N., MEISENZAHL E. M., DAVATZIKOS C., BOTTLENDER R., FRODL T., SCHEUERECKER J., SCHMITT G., ZETZSCHE T., DECKER P., REISER M., ET AL.: Use of neuroanatomical pattern classification to identify subjects in at-risk mental states of psychosis and predict disease transition. *Archives of general psychiatry 66*, 7 (2009), 700–712. 1

[KSC*08] KLÖPPEL S., STONNINGTON C. M., CHU C., DRAGANSKI B., SCAHILL R. I., ROHRER J. D., FOX N. C., JACK JR C. R., ASHBURNER J., FRACKOWIAK R. S.: Automatic classification of mr scans in alzheimer's disease. *Brain 131*, 3 (2008), 681–689. 1

[LPP15] LONG J. D., PAULSEN J. S., PREDICT-HD INVESTIGATORS AND COORDINATORS OF THE HUNTINGTON STUDY GROUP : Multivariate prediction of motor diagnosis in huntington's disease: 12 years of predict-hd. *Movement Disorders 30*, 12 (2015), 1664–1672. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mds.26364, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/mds.26364, doi:10.1002/mds.26364. 6

[MMGP06] MCROBBIE D. W., MOORE E. A., GRAVES M. J., PRINCE M. R.: *MRI from Picture to Proton*, 2 ed. Cambridge University Press, 2006. doi:10.1017/CBO9780511545405. 2

[PC13] PATLE A., CHOUHAN D. S.: Svm kernel functions for classification. In *2013 International Conference on Advances in Technology and Engineering (ICATE)* (Jan 2013), pp. 1–9. doi:10.1109/ICAdTE.2013.6524743. 5

[PHS*14] PLIS S. M., HJELM D. R., SALAKHUTDINOV R., ALLEN E. A., BOCKHOLT H. J., LONG J. D., JOHNSON H. J., PAULSEN J. S., TURNER J. A., CALHOUN V. D.: Deep learning for neuroimaging: a validation study. *Frontiers in Neuroscience 8* (2014), 229. URL: https://www.frontiersin.org/article/10.3389/fnins.2014.00229, doi:10.3389/fnins.2014.00229. 2, 5, 6, 7

[Pol07] POLDRACK R. A.: Region of interest analysis for fMRI. *Social cognitive and affective neuroscience 2*, 1 (mar 2007), 67–70. URL: https://pubmed.ncbi.nlm.nih.gov/18985121https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2555436/, doi:10.1093/scan/nsm006. 5

[PSK02] PASSINGHAM R. E., STEPHAN K. E., KÖTTER R.: The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience 3*, 8 (2002), 606–616. 7

[SKH*08] STEPHAN K. E., KASPER L., HARRISON L. M., DAUNIZEAU J., DEN OUDEN H. E. M., BREAKSPEAR M., FRISTON K. J.: Nonlinear dynamic causal models for fMRI. *NeuroImage 42*, 2 (aug 2008), 649–662. URL: https://pubmed.ncbi.nlm.nih.gov/18565765https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2636907/, doi:10.1016/j.neuroimage.2008.04.262. 2

[SMP*07] STEPHAN K. E., MARSHALL J. C., PENNY W. D., FRISTON K. J., FINK G. R.: Interhemispheric integration of visual processing during task-driven lateralization. *Journal of Neuroscience 27*, 13 (2007), 3512–3522. 7

[WEG87] WOLD S., ESBENSEN K., GELADI P.: Principal component analysis. *Chemometrics and intelligent laboratory systems 2*, 1-3 (1987), 37–52. 6

[YRYR19] YAO H., REGAN M., YANG Y., REN Y.: Image decomposition and classification through a generative model. pp. 400–404. doi:10.1109/ICIP.2019.8802991. 3