Mapping individual differences in cortical architecture using multi-view representation learning

Akrem Sellami [1,2], François-Xavier Dupé [1], Bastien Cagna [2], Hachem Kadri [1], Stéphane Ayache [1], Thierry Artières [1,3], and Sylvain Takerkart [2]

[1]*Aix Marseille Université, Université de Toulon, CNRS, LIS, Marseille, France*
[2]*Institut de Neurosciences de la Timone UMR 7289, Aix-Marseille Université, CNRS, Marseille, France*
[3]*École Centrale de Marseille*
{akrem.sellami, francois-xavier.dupe, hachem.kadri, stephane.ayache, thierry.artieres}@lis-lab.fr
{bastien.cagna, sylvain.takerkart}@univ-amu.fr

***Abstract*—In neuroscience, understanding inter-individual differences has recently emerged as a major challenge, for which functional magnetic resonance imaging (fMRI) has proven invaluable. For this, neuroscientists rely on basic methods such as univariate linear correlations between single brain features and a score that quantifies either the severity of a disease or the subject's performance in a cognitive task. However, to this date, task-fMRI and resting-state fMRI have been exploited separately for this question, because of the lack of methods to effectively combine them. In this paper, we introduce a novel machine learning method which allows combining the activation- and connectivity-based information respectively measured through these two fMRI protocols to identify markers of individual differences in the functional organization of the brain. It combines a multi-view deep autoencoder which is designed to fuse the two fMRI modalities into a joint representation space within which a predictive model is trained to guess a scalar score that characterizes the patient. Our experimental results demonstrate the ability of the proposed method to outperform competitive approaches and to produce interpretable and biologically plausible results.**

***Index Terms*—individual differences, multi-view representation learning, multimodal deep autoencoder, trace regression**

I. Introduction

Since its discovery in the 1990s, functional magnetic resonance imaging (fMRI) has played a critical role in developing our understanding of the brain and its dysfunctions. The research based on fMRI has primarily focused on identifying population-wise principles that emerge from averaging observations across a group of subjects. In cognitive neuroscience, fMRI has greatly furthered our knowledge of the organization of the brain of healthy individuals by mapping *brain activity* induced by a set of tasks that the participant performs in a controlled manner within the scanner (task-fMRI). More recently, the mapping of *functional connectivity* from fMRI recordings of participants lying at rest in the scanner (rest-fMRI) has complemented our understanding of the functional architecture of the brain. In particular, because of its ease of acquisition, rest-fMRI has become a tool of choice in clinical neuroscience to characterize brain dysfunction in groups of patients.

Usually, the specificities of the fMRI signatures extracted in each subject are simply regarded as noise. However, it is now clear that these specificities carry information that can be used to characterize individual differences, measured e.g as the subject's performance in a cognitive task or on a clinical scale that quantifies the severity of disease (see [1], and also the full issue of NeuroImage dedicated to this question [2]). The standard methods used by neuroscientists to study individual differences are based on univariate correlational analysis between a chosen brain feature and such a behavioral score. While it is now clearly established that the design of predictive models should be favoured in this context [3], and that multi-modal data fusion provides a strong added value [4], the question of fusing features extracted from task- and rest-fMRI has not been addressed [4] even though their complementarity is clearly established [5]. In this paper, we introduce a novel machine learning method that exactly aims at combining the activation- and connectivity-based information estimated from different fMRI protocols to build a predictive model of individual differences.

First, the challenge of combining heterogeneous information from task- and rest-fMRI is met for the first time by capitalizing on the recent advances in unsupervised representation learning, and more precisely by developing a specific multi-view deep autoencoder. The main goal of multi-view, or multi-modal, representation learning is to learn a latent space which combines the information from different views available on the same data, where such a latent representation space is assumed to be informative enough to reconstruct the corresponding views [6], [7]. According to the literature, the most common ways of dealing with multiple views consist in either concatenating the view-specific layers (a direct fusion), sharing a hidden layer [8], [9], (e.g. using canonical correlation analysis [10]–[12]) or finding a direct projection from one view to another [13]. In particular, the concatenation of layers can provide efficient representations of multi-modal data when the relationships between modalities are non-linear [8], such as it may be expected with task- and rest-fMRI data.

Secondly, a trace regression model is estimated in this representation space to allow predicting individual characteristics of new patients, measured as their behavioral performance or the severity of their pathology. This also yields a mapping of the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

brain that makes it possible to identify the nodes of the relevant macroscopic cortical network(s). The experimental validation is conducted on a recently published public dataset that aims at studying individual differences in voice recognition abilities using multi-modal MRI data [14]. Our experiments quantitatively demonstrate the advantages provided by the combination of task- and rest-fMRI when compared to an equivalent model learnt on single-view representations. They also show that our multi-view representation learning frawework yields interpretable results that are consistent with the most recent neuroscientific literature.

The remainder of the paper is organized as follows. In Section II, we detail the proposed approach, including the multi-view autoencoder that allows fusing task- and rest-fMRI data, and the trace regression predictive model. In Section III, we first describe the dataset itself together with the preprocessing pipeline that yields multi-modal features; then we detail our experimental protocol, a quantitative evaluation that demonstrates the superiority of the multi-view autoencoder and a neuroscientific interpretation that shows the applicative potential of our method. We conclude in Section IV by discussing opportunities for future work.

II. Method

In this section, we detail the different methodological components of the proposed approach. The goal of this study is to design a model able to predict the behavioral score of a subject (e.g. the performance in a voice recognition task) from multi-view fMRI data of this subject. The main aspects of this setting is the multimodal and high dimensional nature of the raw data and the very small number of subjects (usually tens).

To tackle such a challenge, we choose to design a multimodal feature extractor that combines task- and rest-fMRI point wise (it is learned on isolated vertices of the cortical surface) in order to get enough training material. The fusion of the complementary activation- and connectivity-based information is performed by a deep multi-view autoencoder operating on input pairs of low dimensional feature vectors representing both fMRI modalities for a given vertex. This model is learned on hundreds of thousands of data points.

This feature extractor is used to project the whole brain of each subject into a new representation space within which a predictive model of inter-individual differences (trained to predict a behaviorial score) is learned using a trace regression model estimated in the matrix-valued space generated by the bottleneck layer of the autoencoder. This predictive model operates at the subject level and is then learned on tens of samples only. Figure 1 provides a general overview of this approach.

A. Multimodal deep autoencoder for learning brain features representation

We designed a multimodal autoencoder that aims at extracting high level features from the combination of multiple input modalities, from which these multiple input modalities may be reconstructed [8].

We investigated two main autoencoder architectures which are illustrated in Figure 2. In both cases, the input to the autoencoder is a pair (x_r, x_t) of the two MRI modalities (inputs), x_r for rest-fMRI and x_t for task-based fMRI.

A first architecture, hereafter simply called autoencoder (and noted AE) consists in a standard autoencoder which takes as input the concatenation of inputs $x = concat(x_r, x_t)$. The AE includes an encoder noted E_{AE} and a decoder which are both implemented as dense multi-layer neural networks (with one to three hidden layers). The encoder non-linearly projects the concatenated inputs into a new representation space $z = E_{AE}(x)$ from which the decoder D_{AE} aims to recover x, i.e. $D_{AE}(E_{AE}(x)) \approx x$. The autoencoder is classically trained using a Mean Squared Error (MSE) criterion:

$$\frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \| D_{AE}(E_{AE}(x^{i,j})) - x^{i,j} \|^2$$

where m is the number of vertex of the 3D cortical mesh (see hereafter) and n is the number of subjects, and $x^{i,j} = concat(x_r^{i,j}, x_s^{i,j})$ is the $(i,j)^{th}$ training sample corresponding to the concatenation of the two fMRI data feature vectors for a given vertex i and for subject m, $x_r^{i,j}$ and $x_t^{i,j}$.

Alternatively, we investigated a multi-view deep autoencoder (MDAE) which includes one specific encoder per modality (view), noted E_r and E_t for rest-fMRI and task-based fMRI inputs. The two encoders project each of the modalities into a new representation space $z_r = E_r(x_r)$ and $z_t = E_t(x_t)$ where the representations are concatenated ($z = concat(z_r, z_t)$) and input to two decoders D_r and D_t that aim at recovering the corresponding inputs x_r and x_t. Here again all encoders and decoders are implemented as multi-layer neural networks (we used one to three hidden layers) and the training is performed by minimizing a MSE criterion:

$$\frac{1}{n \times m} \sum_{i=1}^{n} \sum_{j=1}^{m} \| D_r(concat(E_r(x_r^{i,j}), E_t(x_t^{i,j}))) - x_r^{i,j} \|^2$$
$$+ \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} \| D_t(concat(E_r(x_r^{i,j}), E_t(x_t^{i,j}))) - x_t^{i,j} \|^2$$

After training, the models estimated using these two architectures may then be used to compute a multi-view encoding of an input pair (x_r, x_t) through $z_{AE} = E_{AE}(concat(x_r, x_t))$ or $z_{MDAE} = concat(E_r(x_r), E_t(x_t))$.

B. Predictive model of individual differences

We now aim at building a model that is able to predict a scalar score associated with each subject, e.g a behavioral score that measures the performance of the subject in a cognitive task, from the multi-view encoding z available at all cortical locations / vertices. We therefore frame this problem as a regression task. Linear regression has a long history [15], from the least-squares algorithm to the Lasso and its extensions. Modern methods combine common regression loss functions with appropriate regularization such as elastic net [16], (group)-sparsity, manifold [17]. Sparsity constraints

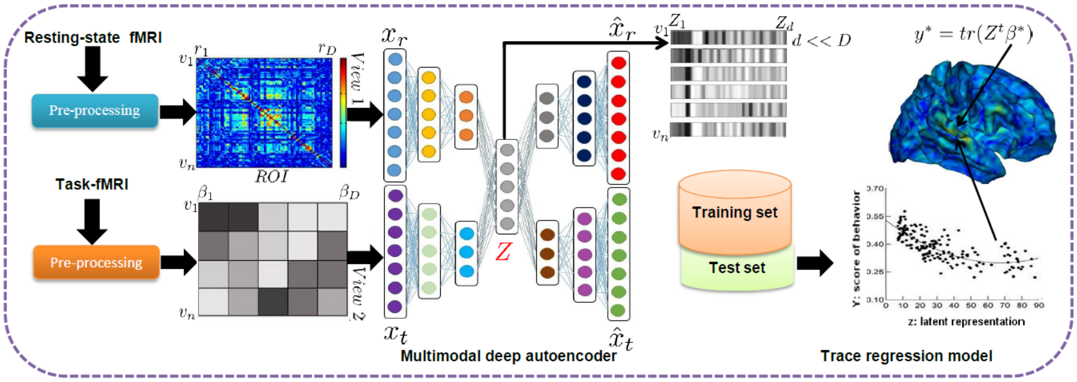

Fig. 1. General overview of the proposed machine learning method

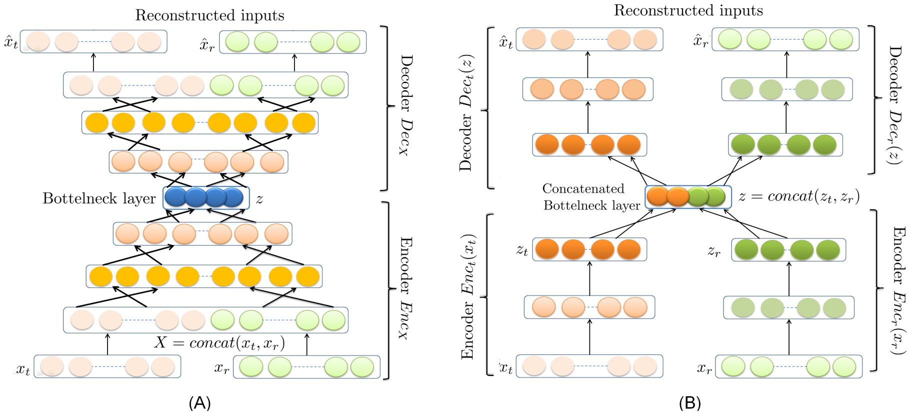

Fig. 2. The two architectures considered for the multimodal representation learning scheme. (A) AE, a standard autoencoder trained on the concatenation of inputs from different modalities. (b) MDAE, a multimodal deep autoencoder where latent representations from the different modalities are concatenated in the bottleneck layer (z is the combination of z_r and z_t).

allow dealing with small sample sizes by seeking for a low dimensional linear subspace. Manifold constraints enforce the solution to belong to a given topology (a mesh for example) encoded by a graph Laplacian. While classical linear regression models seek to predict a scalar response from a vector-valued input, several extensions have been proposed to handle structured inputs such as matrices or tensors [18]. In this case, regularization based on low-rankness or group sparsity are used to capture the structure of the data [18], [19].

In neuroimaging, previous studies (e.g. [20]) approach the behavioral score prediction problem by learning a standard linear regression model using a training set $\{(z_i, y_i)\}$, $i = 1, \ldots, n$, where $z_i \in \mathbb{R}^d$ is a vector of explanatory variables, $y_i \in \mathbb{R}$ is the output response (i.e., the behavioral score), and n is the number of examples (subjects). In contrast, our approach exploits the specific characteristics of our data. First the multi-view features computed by the deep autoencoder form a matrix of explanatory variables $Z_i \in \mathbb{R}^{m \times d}$, where m is the number of vertices of each subject and d is the number of extracted features (size of the multi-view encoding). In order to handle such matrix-valued inputs, we here consider a *trace regression* model. Trace regression is a natural model for matrix-type input data [18], [21], [22]. It provides a linear model that maps matrices to real-valued outputs, and is a generalization of the well-studied linear regression model. The trace regression model corresponds to re-arranging the coefficients in the traditional linear model as a matrix. It is defined as follows

$$y_i = \mathrm{tr}(\beta^{\star\top} Z_i) + \epsilon_i, \;\; i = 1, \ldots, n,$$

where $\mathrm{tr}(\cdot)$ denotes the trace, $Z_i \in \mathbb{R}^{m \times d}$ is a matrix-variate input, $\beta^\star \in \mathbb{R}^{m \times d}$ is the matrix of regression coefficients, y_i is the scalar output and ϵ_i is a random noise.

Moreover, we consider a brain mesh as a graph for designing a specific regularization. Most of previous studies on trace regression perform estimation of the model parameters via regularized least squares with a matrix norm regularization:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{m \times d}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \mathrm{tr}(\beta^\top Z_i) \right)^2 + \Omega(\beta) \right\}, \qquad (1)$$

where $\Omega(\beta)$ is a matrix regularizer which may lead to low-rank or/and structured sparse solutions [18], [19], [22]. In our case, we want to exploit the knowledge that, $\forall i, Z_i$ are related to vector on a mesh (the brain surface), and it should give some benefit to link $\hat{\beta}$ with the underlying manifold. Following [20], we use a manifold regularization which forces each vertex to have feature values close to each other. Formally, given L the Laplacian of the graph of the underlying manifold, we define the regularization term as: $\Omega_1(\beta) = \eta \mathrm{tr}(\beta^\top L \beta)$. Furthermore, since we want to obtain interpretable weights β, we add a sparse prior on the vertices such as only few vertices should be involved in the regression. This is typically done via a suitable group-sparsity regularization strategy. We then consider a second regularization term $\Omega_2(\beta) = \alpha \sum_j \|\beta_j\|^2$, where β_j are the coefficients of the vertex j. Therefore, our predictive model is based on solving the trace regression problem (1) with the regularization term

$$\Omega(\beta) = \eta \mathrm{tr}(\beta^\top L \beta)/2 + \alpha \sum_j \|\beta_j\|^2 \ . \qquad (2)$$

Since $\Omega(\cdot)$ is a convex but non-differentiable function, we use the FISTA algorithm, an accelerated projected gradient descent method, to solve problem (1). We use the monotonous version for stability purpose [23].

Algorithm 1 The main pipeline

```
Input: $\{x_t^i\}_{i=1}^n \in \mathbb{R}^{n \times D_{task}}$, $\{x_r^i\}_{i=1}^n \in \mathbb{R}^{n \times D_{rest}}$: a database of two-view observations, $Y \in \mathbb{R}^n$: the regression target.
    1. Learn a multi-view latent representation $z$ from $x_r$ and $x_t$.
    2. From the latent representation built the resulting observation, $\{z_i\}_{i=1}^n \in \mathbb{R}^{n_z \times D}$.
    3. Apply the trace regression model using $\{Z_i\}_{i=1}^n \in \mathbb{R}^{n_z \times D}$ and $y$ as inputs.
Output: $\beta \in \mathbb{R}^{n_z \times D}$ map used for prediction.
```

III. Experiments

In this section, we describe the experiments that we performed to assess the relevance and effectiveness of our method. We first describe the fMRI dataset that was used, before presenting a thorough quantitative evaluation of both the multi-view representation learning phase and the predictive model itself. Finally, we discuss the neuroscientific relevance of our results. The pipeline used in this section is summarized by Algorithm 1.

A. Data set and pre-processing

We performed our experiments on the InterTVA dataset [14], which aims at studying inter-individual differences in voice perception and voice recognition using multi-modal MRI data. We targeted this data set because it offers both task- and rest-fMRI data, as well as a precise behavioral characterization of the 40 subjects available. More precisely, we exploited: i) the task-fMRI session during which the participants listened passively to 144 vocal and non-vocal sounds, ii) the 12mn long rest-fMRI session, iii) the performance of the subjects in a voice recognition task, measured as percentage the Glasgow Voice Memory Test (GVMT) [24], as well as iv) the high-resolution T1 anatomical image for pre-processing purposes.

Both the task- and rest-fMRI data sets were first corrected for slice-timing and subject's motion (using SPM12, www.fil.ion.ucl.ac.uk/spm), and the physiological noise components were removed using the Physio Tapas toolbox (www.tnu.ethz.ch/en/software/tapas.html). The corrected fMRI time-series were projected on a triangulated mesh representing the surface of the cortex (comprising $20,484$ vertices), that had been extracted from the T1 MRI image using the freesurfer software suite (surfer.nmr.mgh.harvard.edu). Then, two feature vectors were estimated at each cortical location, i.e. each vertex of the mesh, one for each fMRI protocol. For task-fMRI, the feature vector $x_t \in \mathbb{R}^{D_{task}}$ is the set of amplitudes of the fMRI responses induced by each of the 144 audio stimuli presented to the participant, estimated using a general linear model that included one regressor per stimulus [25] (hence, $D_{task} = 144$). For rest-fMRI, the feature vector $x_r \in \mathbb{R}^{D_{rest}}$ contains the correlation coefficients between the time series of the vertex and each of the 150 average time-series computed in the regions of the Destrieux atlas, available in freesurfer (hence, $D_{rest} = 150$). These two feature vectors are then used as inputs of the multi-view representation learning algorithms described in Section II.A. With a training set size of 36 subjects out of the 40 available (see hereafter), the multimodal autoencoder was trained with $36 \times 20,484$ samples of dimension $D_{concat} = D_{rest} + D_{task} = 294$. The task addressed in our experiments was to predict the GVMT score of each participant from the full-brain information carried in these connectivity and activation feature sets, using the trace regression model trained on 36 subjects.

B. Deep autoencoders implementations

We used different types of autoencoders: linear and non linear monomodal (simple) autoencoder (AE), multimodal deep autoencoder (MDAE). For all the models, we investigated several encoding dimensions from 2 to 100 ($enc \in [2,..,100]$). Table I reports the different architectures of AE and MDAE models.

For all the network models, we tested different pairs of activation functions for the hidden layers and output layer, respectively: (*linear, linear*), (*linear, sigmoid*), (*relu, linear*), and (*relu, sigmoid*).

All different autoencoders models were implemented using the *keras* toolkit. Training was performed using the Adam optimizer [26], with a learning rate (lr) equals to 10^{-3} over 300 epochs and a batch size of 500 samples.

C. Benchmarked representation learning methods

We performed a comparative study between several representation learning models, including the two MDAE models (concatenated inputs, and concatenated latent representations), standard AE, principal component analysis (PCA) [27], independent component analysis (ICA) [28], deep canonically correlated autoencoders (DCCAE) [10].

TABLE I

DETAILS ON ARCHITECTURES OF INVESTIGATED AE/MDAE MODELS. THE MODELS ARE BUILT BY STACKING FULLY CONNECTED LAYERS WHOSE DIMENSIONS ARE GIVEN FROM INPUT-LAYER TO OUTPUT LAYER, WITH *enc* BEING THE DIMENSION OF THE ENCODING. FOR INSTANCE $[D_{concat}, enc, D_{concat}]$ STANDS FOR A ONE HIDDEN LAYER ARCHITECTURE TAKING AS INPUT (AND LEARNED TO OUTPUT) THE CONCATENATION OF x_r AND x_t, WHILE PROJECTING THROUGH A HIDDEN LAYER OF SIZE *enc*.

Model (Number of hidden layers)	Architecture
AE (one layer)	$[D_{task}, enc, D_{task}]$ $[D_{rest}, enc, D_{rest}]$ $[D_{concat}, enc, D_{concat}]$
MDAE/AE (two layers)	$[D_{task}, 120, enc, 120, D_{task}]$ $[D_{task}, 130, enc, 130, D_{task}]$ $[D_{rest}, 120, enc, 120, D_{rest}]$ $[D_{rest}, 130, enc, 130, D_{rest}]$ $[D_{concat}, 150, enc, 150, D_{concat}]$ $[D_{concat}, 200, enc, 200, D_{concat}]$
MDAE/AE (three layers)	$[D_{task}, 140, 120, enc, 120, 140, D_{task}]$ $[D_{task}, 140, 130, enc, 130, 140, D_{task}]$ $[D_{rest}, 140, 120, enc, 120, 140, D_{rest}]$ $[D_{rest}, 140, 130, enc, 130, 140, D_{rest}]$ $[D_{concat}, 250, 150, enc, 150, 250, D_{concat}]$ $[D_{concat}, 200, 130, enc, 130, 200, D_{concat}]$

1) PCA: as baseline, we investigated the use of PCA to reduce the dimensionality of multimodal fMRI data. PCA seeks an optimal linear orthogonal transformation that provides a new coordinate system, i.e., the latent space, in which basis vectors follow modes of greatest variance in the original fMRI data [27].

2) ICA: an extension of PCA technique, which aims to optimize higher-order statistics such as kurtosis. Usually, it is used as a computational method for separating a multivariate signals into additive sub-components [28].

3) DCCAE: the aim is to optimize the combination of canonical correlation between the learned bottleneck representations (latent space) and the reconstruction errors of the autoencoders [10]. In fact, it consists of two autoencoders on top of two deep neural networks.

D. Evaluation scheme

In order to train the different models, we perform a 10-fold cross-validation, i.e., for each fold, we used 36 subjects for training and 4 subjects for testing. Moreover, we compared all representation learning methods (AE, MDAE, DCCAE, PCA, and ICA) based on the prediction error by computing the MSE between the true behavioral score noted by y and the predicted behavioral score estimated by the trace regression model noted by $\hat{y}$. Therefore, we tested all methods on 10-fold cross-validation, and we reported the average MSE for each method. Moreover, we computed the average R-squared (coefficient of determination) noted by R^2, which is the fraction by which the variance of the errors is less than the variance of the dependent variable. It can be defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y}_i)^2} \tag{3}$$

where $\hat{y}_i$ is the predicted behavioral score and the $\bar{y}_i$ is the mean of behavioral scores. The R^2 indicates how well the model predictions approximate the true values.

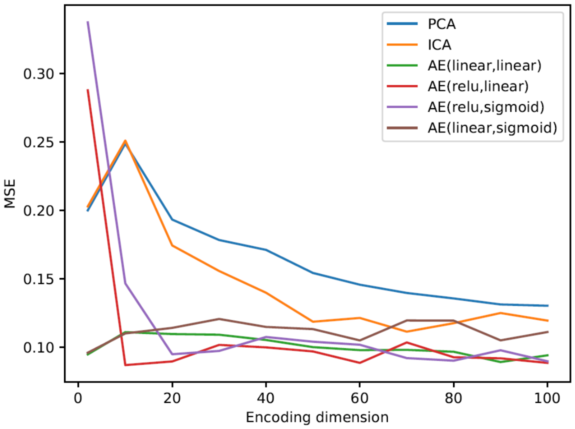

Fig. 3. Average MSE versus encoding dimension across 10-fold cross validation using task-based fMRI data

E. Results: quantitative evaluation of the predictive model

In this section, we present the results obtained by our predictive model. The trace regression model (II-B) is estimated on the latent representation $Z \in \mathbb{R}^{n \times d}$. For the spatial constraints, we use the Laplacian matrix of the graph given by the triangulated cortical mesh. Indeed, we use the null vector as initialization in order to catch only useful features. This spatial regularization makes it possible to favor the nodes having the same value as their neighbors. We empirically found by cross-validation α and η respectively equals to $5e^{-4}$ and $1e^{-3}$ are the best hyper-parameters values. Moreover, in order to benchmark our approach, we also applied the trace regression model on the raw fMRI data. Hence, we perform two main experimentation for the evaluation of the predictive model, using unimodal and multimodal fMRI data.

1) Prediction performances using monomodal fMRI data: In order to study individual differences from such a dataset, a neuroscientist would classically use the single data modality that relates the most closely to nature of the task used to measure the behavioral performance, hence task-fMRI (listening to voices in task-fMRI, recognizing voices in the GVMT behavioral score). We therefore provide a comparative study between representation learning methods, including PCA, ICA, linear AE, and non-linear AE with different pairs activation functions using monomodal task-fMRI data. The aim, is to study the influence of the nature of the representation learning models on the regression task. Figure 3 reports the average MSE versus encoding dimension using task-fMRI data. We can notice that the best MSE for task-fMRI is obtained using the AE*(relu, linear)* model, which is equal to 0.09, with an encoding dimension of 10. We can also observe that all AE models are effective with few dimensions (*enc* is between 10 and 20 features), while the other models require using many more dimensions to reach their best performance, which demonstrates the efficiency of the AE models at compressing the

TABLE II

Best average MSE and R^2 ($\pm$ standard error) using monomodal fMRI data estimated on trace regression model after regularization with spatial constraint ($L1$) and parsimony ($L2$) based on PCA, ICA, and different AE models (3 layers is the best architecture: [D_{task}, 140, 120, enc, 120, 140, D_{task}] and [D_{rest}, 140, 120, enc, 120, 140, D_{rest}]).

	task-fMRI		rest-fMRI	
	Average MSE	Average R^2	Average MSE	Average R^2
PCA	0.152 ($\pm$ 0.0035)	0.129 ($\pm$ 0.0051)	0.345 ($\pm$ 0.0082)	0.071 ($\pm$ 0.0047)
ICA	0.137 ($\pm$ 0.0094)	0.162 ($\pm$ 0.0085)	0.345 ($\pm$ 0.0032)	0.073 ($\pm$ 0.0124)
AE (*linear, linear*)	0.091 ($\pm$ 0.0071)	0.239 ($\pm$ 0.0023)	**0.340** ($\pm$ 0.0195)	**0.094** ($\pm$ 0.0025)
AE (*linear, sigmoid*)	0.093 ($\pm$ 0.0026)	0.201 ($\pm$ 0.0096)	0.344 ($\pm$ 0.0088)	0.081 ($\pm$ 0.0040)
AE (*relu, linear*)	**0.082** ($\pm$ 0.0088)	**0.259** ($\pm$ 0.0075)	0.344 ($\pm$ 0.0163)	0.086 ($\pm$ 0.0019)
AE (*relu, sigmoid*)	0.102 ($\pm$ 0.0053)	0.112 ($\pm$ 0.0029)	0.342 ($\pm$ 0.0179)	0.089 ($\pm$ 0.0201)
Raw fMRI data	0.632 ($\pm$ 0.0289)	-2.703 ($\pm$ 0.3055)	0.905 ($\pm$ 0.0179)	-8.384 ($\pm$ 0.4241)

TABLE III

Best average MSE and R^2 ($\pm$ standard error) using multimodal fMRI data estimated on trace regression model after regularization with spatial constraint ($L1$) and parsimony ($L2$) based on PCA, ICA, DCCAE, and different AE/MDAE models (3 layers is the best architecture: [D_{concat}, 200, 130, enc, 130, 200, D_{concat}]).

	Concatenated inputs (task+rest-fMRI)		Concatenated latent representations ($z_t + z_r$)	
	Average MSE	Average R^2	Average MSE	Average R^2
PCA	0.091 ($\pm$ 0.0021)	0.239 ($\pm$ 0.0048)	N/A	N/A
ICA	0.092 ($\pm$ 0.0109)	0.223 ($\pm$ 0.0090)	N/A	N/A
AE/MDAE (*linear, linear*)	0.083 ($\pm$ 0.0067)	0.254 ($\pm$ 0.0104)	0.065 ($\pm$ 0.0058)	0.273 ($\pm$ 0.0033)
AE/MDAE (*linear, sigmoid*)	0.083 ($\pm$ 0.0089)	0.249 ($\pm$ 0.0036)	0.079 ($\pm$ 0.0124)	0.249 ($\pm$ 0.0204)
AE/MDAE (*relu, linear*)	**0.080** ($\pm$ 0.0032)	**0.262** ($\pm$ 0.0136)	**0.062** ($\pm$ 0.0095)	**0.282** ($\pm$ 0.0044)
AE/MDAE (*relu, sigmoid*)	**0.080** ($\pm$ 0.0074)	0.260 ($\pm$ 0.0281)	0.073 ($\pm$ 0.0086)	0.254 ($\pm$ 0.0051)
DCCAE	N/A	N/A	0.126 ($\pm$0.0075)	0.183 ($\pm$0.0093)
Raw fMRI data ($x_t + x_r$)	N/A	N/A	-	-

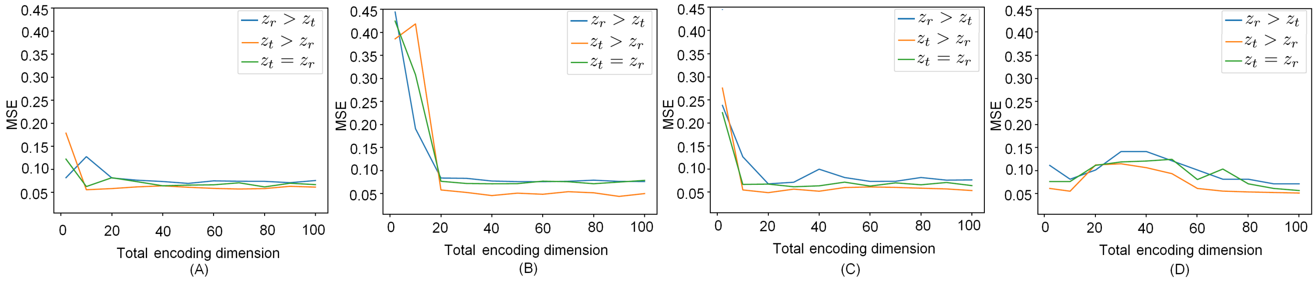

Fig. 4. Comparison of the average MSE (over the 10-fold cross-validation) for different relative sizes given to the task- and rest-fMRI modalities in the bottleneck layer of the MDAE (larger size for rest-fMRI: $z_r > z_t$; larger size for task-fMRI: $z_t > z_r$; equal sizes: $z_t = z_r$ (A). MDAE (*linear, linear*), (B). MDAE (*linear, sigmoid*), (C). MDAE (*relu, linear*), (D). MDAE (*relu, sigmoid*). Overall, the performances are higher when $z_t > z_r$.

data efficiently in this context. For the sake of exhaustivity, we also compare the performances of the models estimated from monomodal task-fMRI data and from monomodal rest-fMRI data: Table II includes the best (across encoding dimension) average MSE and R^2 over the 10-fold cross validation, for all representation learning models. The performances obtained from monomodal rest-fMRI data are far worst than the ones obtained for monomodal task-fMRI data. This was to be expected since the task-fMRI data involved a cognitive task that was very close to the one performed during the behavioral GVMT test, while the rest-fMRI data only contains non-specific information of the whole-brain functional connectivity. Moreover, the performances were also a lot better when using a representation learning method than when using the raw data (last row of Table II), both in terms of MSE and R^2. This demonstrates the added value of using such representation learning approach for this task.

2) Multimodal representation learning: combining task- and rest-fMRI: In this section, we compare the different implementations of the two multimodal autoencoder introduced in Section II.A, i.e. when the task- and rest-fMRI inputs are combined to learn a multi-view representation. We compare them, again, using MSE and R^2. Hence, we report in Table III the best (across encoding dimensions) average MSE and R^2 over 10-fold cross validation using combined fMRI data for each method. The best performances obtained using the first architecture (AE, concatenation of the inputs, left part of Table III) are MSE = 0.080 and $R^2 = 0.262$ for the AE(*relu, linear*) model, which are barely better than the performances obtained from the monomodal model based on task-fMRI. However, we obtain an important gain of performances with our second multi-view architecture (MDAE, concatenation

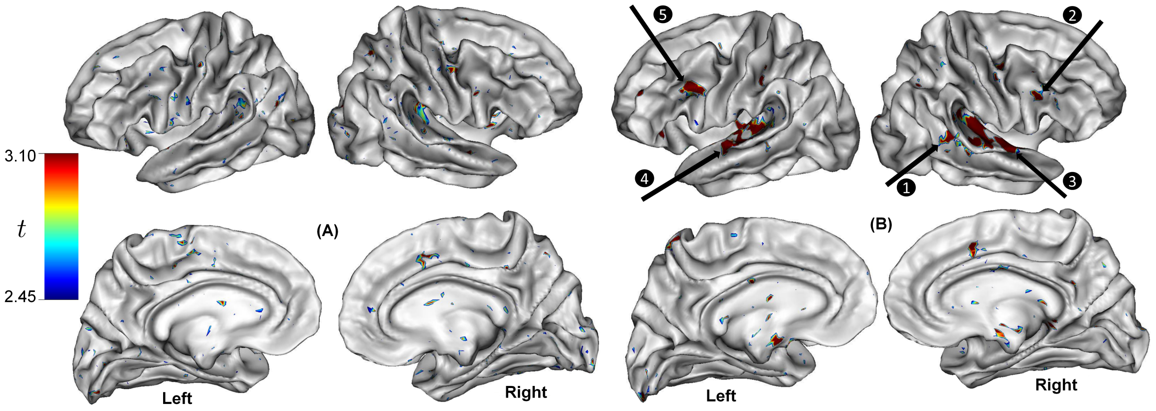

Fig. 5. Average weight maps estimated using our task+rest fMRI-based predictive model of individual differences, thresholded after a test for statistical significance ($t > 2.45$, $p < 0.005$). Regions of non zero average weight appear in color. (A) With the model using concatenated inputs in the MDAE (*relu, sigmoid*), only small scattered regions are detected. (B) When using the concatenated latent representations in the MDAE (*relu, sigmoid*), several larger regions are detected: regions tagged 3 and 4 are along the superior temporal gyrus bilaterally, region 1 is located in the fundus of the right superior temporal sulcus, regions 2 and 5 are in the inferior frontal gyrus. All these regions perfectly match the neuroscientific literature [29]–[31].

of the latent representations): all MDAE models actually outperform the monomodal models, the multimodal models based on concatenated inputs, as well as the deep canonically correlated autoencoder multimodal model (DCCAE). The best model is the MDAE(*relu, linear*) model, with a MSE = 0.062 and $R^2 = 0.282$. This clearly demonstrates the added value of fusing task- and rest-fMRI in order to study inter-individual differences, and that a multimodal autoencoder that operates a concatenation of the latent representation is an efficient representation learning scheme for this objective.

3) Influence of the relative size of the rest- and task-fMRI in the bottleneck layer of the MDAE: Having demonstrated the effectiveness of our approach, the objective of this section is to attempt to differentially weight the two modalities in order to further improve the performances obtained using our MDAE with concatenated latent representations. To do so, we vary the sizes of the latent representations z_t and z_r of each modality for a given total encoding dimension. We present three configurations where each configuration corresponds to different sizes of the bottleneck layers z_t and z_r. Thereby, in the first configuration, the size of $z_r > z_t$. In the second configuration, the size of $z_r < z_t$, and in the last configuration the sizes of z_r and z_t are equal ($z_r = z_t$). Figure 4 reports the average MSE computed over the 10-fold cross-validation using the four MDAE models with different pairs of activation functions, which are the four best models identified previously (see Table III). Using these three configurations, we aim here at evaluating the importance of each modality for the regression task. Overall, the best MSEs are obtained for the second configuration, i.e. when $z_t > z_r$, showing that the task-fMRI provides more information than the rest-fMRI data. For instance, the best MSE is equal to 0.050 with MDAE(*relu, sigmoid*) and the R^2 is of 0.294 (± 0.0048), where 8 features extracted from task-fMRI and 2 features extracted from rest-fMRI. Furthermore, for the MDAE(*linear, linear*) model, the best average MSE and R^2 are obtained with an encoding dimension of 10, where again $z_t > z_r$. For the MDAE(*relu, linear*) model, 20 features allow to get a higher prediction with 15 features from task-fMRI and 5 features from rest-fMRI. For the MDAE (*linear, sigmoid*) model, the best MSE is equal to 0.052 where 12 features extracted from task-fMRI and 8 features from rest-fMRI (20 features). We can conclude then that the task-fMRI contributes more than the rest-fMRI, where the MSE is very low and the prediction accuracy is sufficient. Moreover, we have demonstrated that the concatenation of bottleneck layer is more suitable than the concatenation of inputs.

F. Neuroscientific relevance of the model

We now examine the neuroscientific relevance of our results. Lacking a ground truth, we compare our results qualitatively with state of the art knowledge extracted from the neuroscientific literature [29]–[31]: knowing the task performed by the subject (i.e the passive listening of vocal sounds), we can expect to see a very focal network of brain regions located bilaterally in the temporal lobe (along the superior temporal gyrus and sulcus [29], [30]), as well as regions in the frontal lobe (in the pre-central and inferior frontal gyri [31]). For this, we present on Fig. 5 the average $\hat{\beta}$ weight maps estimated using the trace regression, for the multi-modal representations obtained with either concatenated inputs or concatenated latent representations. To ease the interpretation, these maps are presented overlayed on the three-dimensional cortical mesh, after performing a statistical test to extract significant clusters of non-zero average weights (t-test, performed in SPM12; thresholded at $p < 0.005$, i.e $t > 2.45$). The weight maps estimated using the concatenation of the inputs (on the left) present very few regions of small sizes where the average weight is significantly non zero, scattered all over the cortex. In contrast, the multi-view autoencoder based on concatenated latent representations (on the right) yields several larger significant regions which closely correspond to the expectations described previously (see arrows on Fig. 5). Remarkably, it allows detecting significant regions in the frontal lobe bilaterally, regions that are known to be hard to detect (e.g. 92 subjects were used in [31] whereas only 39 were available in the present study). This might reflect a gain in statistical power that could be induced by an improved robustness of the information present in the latent representation of the fused task- and rest-fMRI data. Further experimental validation on other datasets should be conducted to confirm this potential gain in statistical power.

IV. Conclusion

In this paper, we introduced a novel machine learning method which aims at mapping individual differences in cortical architecture using multi-view representation learning. Our method seeks to fuse task- and rest-fMRI in order to exploit the complementary information they offer about brain activation and connectivity respectively. To do so, a deep multi-view autoencoder was designed to fuse the two fMRI modalities, yielding a compressed joint representation space within which a trace regression model is developed. This model allows predicting the behavioral performances of new patients. Our experimental results demonstrate the ability of the proposed method to outperform competitive approaches and produce interpretable and biologically plausible results with a potential gain in statistical power. In the future, the proposed method can be extended by introducing a graph convolutional network model incorporating the 3-D cortical mesh in order to preserve the spatial structure, but also by adding extra data modalities that can characterize structural connectivity or the local folding pattern of the cortex.

Acknowledgment

This work was carried out within the *Institut Convergence ILCB* (ANR-16-CONV-0002) and with access to the high performance computing resources of the *Centre de Calcul Intensif d'Aix-Marseille*. It was also funded in part by the French *Agence Nationale de la Recherche* (grants ANR-15-CE23-0026 and ANR-16-CE23-0006).

References

[1] M. L. Seghier and C. J. Price, "Interpreting and Utilising Intersubject Variability in Brain Function," *Trends in Cognitive Sciences*, vol. 22, no. 6, pp. 517–530, Jun. 2018.

[2] V. D. Calhoun, S. M. Lawrie, J. Mourao-Miranda, and K. E. Stephan, "Prediction of individual differences from neuroimaging data," *NeuroImage*, vol. 145, no. Pt B, p. 135, 2017.

[3] J. Dubois and R. Adolphs, "Building a Science of Individual Differences from fMRI," *Trends in Cognitive Sciences*, vol. 20, no. 6, pp. 425–443, Jun. 2016.

[4] V. D. Calhoun and J. Sui, "Multimodal Fusion of Brain Imaging Data: A Key to Finding the Missing Link(s) in Complex Mental Illness," *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, vol. 1, no. 3, pp. 230–244, May 2016.

[5] M. Cole, D. Bassett, J. Power, T. Braver, and S. Petersen, "Intrinsic and Task-Evoked Network Architectures of the Human Brain," *Neuron*, vol. 83, no. 1, pp. 238–251, Jul. 2014.

[6] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[7] Y. Li, M. Yang, and Z. M. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[8] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th international conference on machine learning (ICML-11)*, 2011, pp. 689–696.

[9] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 7–16.

[10] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[11] M. Kan, S. Shan, and X. Chen, "Multi-view deep network for cross-view classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4847–4855.

[12] K. Somandepalli, N. Kumar, R. Travadi, and S. Narayanan, "Multimodal representation learning using deep multiset canonical correlation," *arXiv preprint arXiv:1904.01775*, 2019.

[13] T.-M. H. Hsu, W.-H. Weng, W. Boag, M. McDermott, and P. Szolovits, "Unsupervised multimodal representation learning across medical images and reports," *arXiv preprint arXiv:1811.08615*, 2018.

[14] A. Virginia, C. Bastien, B. Pascal, and T. Sylvain, "A multimodal MRI dataset for the study of inter-individual differences in voice perception and identification," https://openneuro.org/datasets/ds001771/versions/1.0.0, 2019.

[15] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.

[16] Z. Zhang, Z. Lai, Y. Xu, L. Shao, J. Wu, and G.-S. Xie, "Discriminative elastic-net regularized linear regression," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1466–1481, 2017.

[17] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *Journal of machine learning research*, vol. 7, no. Nov, pp. 2399–2434, 2006.

[18] V. Koltchinskii, K. Lounici, and A. B. Tsybakov, "Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion," *The Annals of Statistics*, vol. 39, no. 5, pp. 2302–2329, 2011.

[19] J. Zhao, L. Niu, and S. Zhan, "Trace regression model with simultaneously low rank and row (column) sparse parameter," *Computational Statistics & Data Analysis*, vol. 116, pp. 1–18, 2017.

[20] L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor, "Interpretable whole-brain prediction analysis with graphnet," *NeuroImage*, vol. 72, pp. 304–321, 2013.

[21] M. Slawski, P. Li, and M. Hein, "Regularization-free estimation in trace regression with symmetric positive semidefinite matrices," in *Advances in Neural Information Processing Systems*, 2015, pp. 2782–2790.

[22] J. Fan, W. Gong, and Z. Zhu, "Generalized high-dimensional trace regression via nuclear norm regularization," *Journal of Econometrics*, 2019.

[23] A. Beck and M. Teboulle, "Gradient-based algorithms with applications to signal recovery," *Convex optimization in signal processing and communications*, pp. 42–88, 2009.

[24] V. Aglieri, R. Watson, C. Pernet, M. Latinus, L. Garrido, and P. Belin, "The glasgow voice memory test: Assessing the ability to memorize and recognize unfamiliar voices," *Behavior research methods*, vol. 49, no. 1, pp. 97–110, 2017.

[25] V. Aglieri, B. Cagna, P. Belin, and S. Takerkart, "Single-trial fmri activation maps measured during the intertva event-related voice localizer. a data set ready for inter-subject pattern analysis," *Data in Brief*, p. 105170, 2020.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] C. M. Bishop, *Pattern recognition and machine learning*. springer, 2006.

[28] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[29] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. P ike, "Voice-selective areas in human auditory cortex," *Nature*, vol. 403, no. 6767, pp. 309–312, Jan. 2000, bibtex: belin_voiceselective_2000.

[30] C. R. Pernet, P. McAleer, M. Latinus, K. J. Gorgolewski, I. n. Charest, P. E. Bestelmeyer, R. H. Watson, D. Fleming, F. Crabbe, M. Valdes-Sosa, and P. Belin, "The human voice areas: Spatial organization and inter-individual variability in temporal and extra-te mporal cortices," *NeuroImage*, vol. 119, pp. 164–174, Oct. 2015.

[31] V. Aglieri, T. Chaminade, S. Takerkart, and P. Belin, "Functional connectivity within the voice perception network and its behavioural relevance," *NeuroImage*, vol. 183, pp. 356–365, Dec. 2018.