Deep Multi-View Representation Learning for Multi-modal Features of the Schizophrenia and Schizo-affective Disorder

Jun Qi[1]*, Javier Tejedor*[2]

1. Electrical Engineering, University of Washington, Seattle, USA
2. Department of Electronics, University of Alcala, Madrid, Spain

qij13@u.washington.edu, javier.tejedor@depeca.uah.es

Abstract

This work is originated from the MLSP 2014 Classification Challenge which tries to automatically detect subjects with schizophrenia and schizo-affective disorder by analyzing multi-modal features derived from magnetic resonance imaging (MRI) data. We employ Deep Neural Network (DNN)-based multi-view representation learning for combining multi-modal features. The DNN-based multi-view models include deep canonical correlation analysis (DCCA) and deep canonically correlated auto-encoders (DCCAE). In addition, support vector machine with Gaussian kernel is used to conduct classification with the compact bottleneck features learned by the deep multi-view models. Our experiments on the dataset provided by the MLSP Classification Challenge show that bottleneck features learned via deep multi-view models obtain better results than the trimming features used in the baseline system in terms of the receiver operating characteristic (ROC) area under the curve (AUC).

Index Terms: Deep Canonical Correlation Analysis, Deep Canonically Correlated Auto-encoders, Schizophrenia, MRI, Support Vector Machine, ROC/AUC

1. Introduction

This work is motivated by the MLSP 2014 Classification Challenge [1] which aims to automatically diagnose schizophrenia and schizo-affective disorder (henceforth schizophrenia) by making use of multi-modal features derived from magnetic resonance imaging (MRI) data. Recent studies demonstrate that the schizophrenia is a brain disease that affects more than 0.7% of people across the world [2] and that the schizophrenia can be efficiently detected by analyzing its bio-imaging data [3].

The MLSP 2014 Classification Challenge provides two kinds of multi-modal features derived from MRI data [4, 5]: one is the source-based morphometric loading (SBM) which corresponds to structural magnetic resonance imaging (sMRI) data; the other refers to the functional network connectivity (FNC) which is associated with the resting state functional MRI (rs-fMRI) data. The task for the challenge is to diagnose the schizophrenia or predict the disease onset in subjects who are at risk of psychosis. Although several solutions are proposed for the task, the baseline system chosen by the challenge is based on the simply feature trimming approach followed by the disease classification via support vector machine (SVM), which achieved the second place in the challenge competition [1].

This work, however, aims to enhance the diagnosis of schizophrenia by combining the multi-modal features within deep multi-view representation learning models. Such typical models include 'deep canonical correlation analysis' (DCCA) [6] and 'deep canonically correlated auto-encoders' (DCCAE) [7], both of which are non-linear extensions of the canonical correlation analysis (CCA) [8].

The CCA is a standard technique for unsupervised data analysis which finds linear projections of two random vectors that are maximally correlated. In mathematics, we denote two random vectors (X_1, X_2) with covariance matrices $(\Sigma_{11}, \Sigma_{22})$ and cross-covariance matrix Σ_{12}. The CCA tries to find pairs of linear projections of two views A_1, A_2 that are maximally correlated, as shown in (1):

$$\max_{A_1, A_2} tr(A_1^T \Sigma_{12} A_2) \tag{1}$$

$$s.t., A_1^T \Sigma_{11} A_1 = A_2^T \Sigma_{22} A_2 = I,$$

where I is the identity matrix.

One important application of the CCA relates to learning features for multiple modalities that are then fused for prediction, which is quite fitted to the task of diagnosing the schizophrenia. The limitation of the CCA is that this bases on a linear projection, which makes difficult a compact feature representation. Kernel CCA (KCCA) [9] is an alternative to the CCA since that computes non-linear projections of the two views, as shown in (2):

$$\max_{A_1, A_2} A_1^T K_1 K_2 A_2 \tag{2}$$

$$s.t., A_1^T K_1^2 A_1 = A_2^T K_2^2 A_2 = I,$$

where I is the identity matrix, $K_1, K_2 \in R^{m \times m}$ represent the gram matrices in which $K_1 = K - K\mathbf{1} - \mathbf{1}K + \mathbf{1}K\mathbf{1}$, any entry in K_1 is represented as a kernel function k_1 for two observations x_i, x_j denoted as $K_{ij} = k_1(x_i, x_j)$, $\mathbf{1} \in R^{m \times m}$ is an all-1s matrix, and similarly for K_2.

However, KCCA presents two important drawbacks: (1) the feature representation is limited by the fixed kernel and (2) KCCA is not suitable for large datasets [10].

The most recently proposed techniques for deep multi-view representation learning do not only deliver compact bottleneck features which correspond to some latent patterns shared by two input features, but their computational complexity is also significantly reduced to the extent that they can be scalable to large datasets. The deep multi-view techniques are inspired by the deep neural network (DNN) that allows more than two hidden layers which can be well trained by the layer-by-layer Restricted Boltzmann Machine (RBM) pre-training based on Contrastive divergence [11]. While the DNN has been successfully applied to supervised classification tasks, we use it in an unsupervised way to learn non-linear transformations of two kinds of features to a space in which the data are highly correlated [12].

In this work, DCCA and DCCAE are the two deep multi-view models applied in the diagnosis of schizophrenia.

When bottleneck features have been learned by the deep multi-view models, they are fed to the SVM with Gaussian kernel for a final classification. The SVM follows the same setup as the baseline system, so it is fair to compare the features used in our system with those proposed in the baseline system.

The rest of the paper is organized as follows: Section 2 presents the DNN-based multi-view feature learning models, including the DCCA and the DCCAE. Section 3 introduces our system for the diagnosis of schizophrenia. Experiments are reported in Section 4 and the paper is concluded in Section 5.

2. DNN-based Multi-View Representation Learning

This section presents two DNN-based multi-view feature learning models: the DCCA and the DCCAE. The main difference between both is whether or not the model regularization based on auto-encoders is applied.

2.1. Deep Neural Network

The deep neural network has been widely applied to numerous tasks, such as automatic speech recognition [13], computer vision [14], and natural language processing [15]. The DNN employed in this work allows multi-layer with more than two hidden layers and applies the RBM pre-training in an unsupervised way. The RBMs are trained layer-by-layer and are stacked into a multi-layer perceptron (MLP) when the RBM training is completed.

For classification tasks, a labeled layer is stacked on top of all RBM layers and fine-tunes the parameters of the MLP by back-propagation.

For DCCA and DCCAE unsupervised learning, RBM pre-training for two DNNs is first conducted, and next, the parameters of the DNNs and the linear transformations in the CCA are jointly further learned by the stochastic gradient descent (SGD) aiming at maximizing the correlation between the non-linear transformations of two input features.

When applying unsupervised learning for the DNN auto-encoders, the DNN is stacked by RBMs learned in an unsupervised way, and the shallow layer on top of the DNN reconstructs the inputs with minimum errors.

2.2. Deep canonical correlation analysis (DCCA)

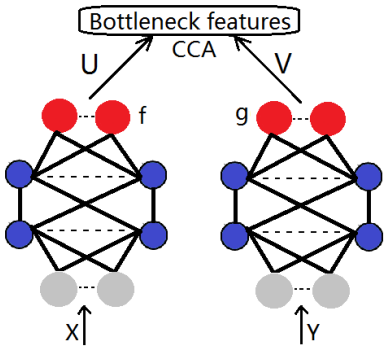

Figure 1: *Deep Canonical Correlation Analysis.*

The DCCA consists of two DNNs and maximizes the canonical correlation of the two DNN outputs, which can be illustrated as Figure 1 and formulated in mathematics as (3):

$$\min_{W_f,W_g,U,V} -\frac{1}{N} tr(U^T f(X) g(Y)^T V) \tag{3}$$

$$s.t., U^T(\frac{1}{N} f(X) f(X)^T + r_X I)U = I$$

$$V^T(\frac{1}{N} g(Y) g(Y)^T + r_Y I)V = I$$

$$u_i^T f(X) g(Y)^T v_j = 0, \quad \forall i \neq j,$$

where I is the identity matrix, N refers to the total number of data, X and Y are two random vectors that represent two input features, $f(\cdot)$ and $g(\cdot)$ represent non-linear transformations of the two DNNs with parameters W_f and W_g respectively, $U = [u_1, ..., u_L]$ and $V = [v_1, ..., v_L]$ refer to the CCA directions that project the DNN outputs to a bottleneck layer with L units, and $(r_x, r_y) > 0$ are regularization parameters for sample covariance estimation. The $U^T f(\cdot)$ is the final non-linear projection mapping used for testing.

As discussed in [6], the DCCA requires all the training data with the whitening constraints, and hence the SGD is not applied in a standard way. However, the DCCA can still be optimized efficiently as long as sufficient large mini-batch data are used for the gradient.

2.3. Deep canonically correlated auto-encoders (DCCAE)

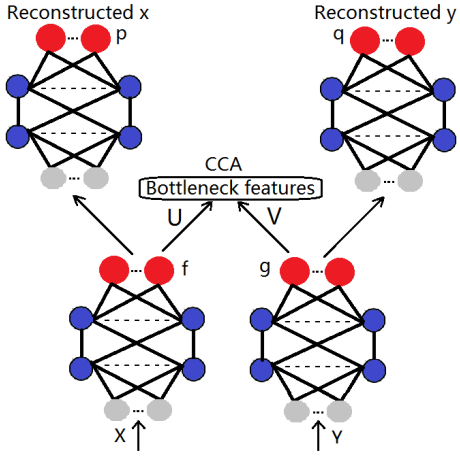

Figure 2: *Deep Canonically Correlated Auto-encoders.*

As shown in Figure 2, the DCCAE consists of two auto-encoders on top of two DNNs and optimizes the canonical correlation of the learned bottleneck features and the reconstruction errors of the auto-encoders. The mathematical formulation for the DCCAE is shown as (4):

$$\min_{W_f,W_g,W_p,W_q,U,V} -\frac{1}{N} tr(U^T f(X) g(Y)^T V) \tag{4}$$

$$+\frac{\lambda}{N} \sum_{i=1}^{N} (||x_i - p(f(x_i))||^2 + ||y_i - q(g(y_i))||^2)$$

$$s.t., U^T(\frac{1}{N}f(X)f(X)^T + r_X I)U = I$$

$$V^T(\frac{1}{N}g(Y)g(Y)^T + r_Y I)V = I$$

$$u_i^T f(X)g(Y)^T v_j = 0, \quad \forall i \neq j,$$

where the symbols I, X, Y, N, W_f, W_g, U, V, r_X, r_Y, and the functions $f(\cdot)$ and $g(\cdot)$ represent the same as in (3), $p(\cdot)$ and $q(\cdot)$ are the non-linear functions associated with the two reconstructed DNNs with parameters W_p and W_q respectively, (x_i, y_i) denotes the i-th data, and λ is a regularization constant that controls the level of the auto-encoders.

Note that the constraints in (4) are the same as those in (3) and that only a regularization term for auto-encoders is added to the DCCA objective.

Similar to the DCCA, the SGD can be applied to the DCCAE objective. The stochastic optimization is the sum of the gradient for the auto-encoders and the gradient for the DCCA.

3. The Diagnosis System

The system for the diagnosis of schizophrenia is shown in Figure 3. The input to the system consists of two multi-modal features that correspond to the SBM and the FNC. The multi-view models are used for the feature representation of the two multi-modal features. The sample classification is conducted by the SVM with Gaussian kernel and outputs a final diagnosis.

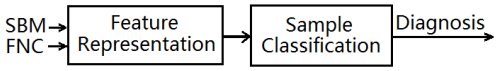

Figure 3: *Diagnosis System for the Schizophrenia.*

Since our work is only concerned with deep multi-view models for feature representation, the sample classification employs the same SVM with Gaussian kernel as the baseline system such that we are able to compare performances obtained by the deep multi-view models and the baseline system. The SVM optimization is shown as (5):

$$\max_{\alpha_1,\ldots,\alpha_N} \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(x_i, x_j, \sigma), \quad (5)$$

where N is the total number of training data, $\{x_r, y_r\}_{r=1}^N$ represents training data, $\{\alpha_r\}_{r=1}^N$ denotes associate dual variables, σ controls the width of the Gaussian kernel, and the Gaussian kernel $K(x_i, x_j, \sigma)$ is defined as (6):

$$K(x_i, x_j, \sigma) = exp(-\frac{||x_i - x_j||^2}{2\sigma^2}). \quad (6)$$

4. Experiments

4.1. Data profile

The database consists of two kinds of multi-modal features, namely the source-based morphometric loading and the functional network connectivity. The SBM and the FNC features are extracted from the structural magnetic resonance imaging data and the resting state functional MRI data respectively. The acquisition and preprocessing details of structural and functional imaging data, including the feature extraction protocol, are introduced in [4]. The number of dimensions for the SBM feature is 32, while there are 378 dimensional FNC features.

The original database contains 86 labeled and 119748 unlabeled data. According to the evaluation requirements of the MLSP 2014 Challenge Competition, 52% of the unlabeled data (i.e., 62269) are specifically selected as test data for the final evaluation, while the rest of unlabeled data (i.e., 57479) are used as training and development for the DNN-based multi-view models. Specifically, 5479 of these are randomly chosen as development set and the rest (i.e., 52000) form the training set. The SVM training is based on the small number of labeled data: first, these labeled data are transformed into compact bottleneck features based on the deep multi-view models, and next, the SVM training is conducted. The diagnosis results for the test data are evaluated by the MLSP 2014 Classification Challenge evaluation system [1].

4.2. Model setup

Two DNNs are set up in the experiments: one is designed for the SBM and the other aims at the FNC. The two DNNs are configured as 4 hidden layers with setups as 256-256-256-256 for the SBM and as 1024-1024-1024-1024 for the FNC. The bottleneck feature dimension is set to 122 such that the features can be equally compared with the 122 dimensional trimming features used in the baseline system. The number of mini-batch data for the DNN training is set to 2048 to ensure enough data are used. The maximum number of iterations and the momentum rate are set to 20 and 0.5 respectively. All these parameters have been tuned on the development set.

As to the particular setup of the DCCAE, the learned bottleneck codes are fed to the hidden layers of the two DNNs in a reverse order such that two reconstructed inputs are obtained. The errors between the original inputs and the reconstructed inputs are then sent again to the top bottleneck layer. The iterations will not terminate until the errors on the development set fall below a given threshold. The regularization constant λ is set to 0.1 for the first 5 iterations and reduced to 0.05 for the rest of the iterations. These values have been optimized on the development set.

4.3. Evaluation metric

The diagnosis results are judged based on the receiver operating characteristic (ROC) area under the curve (AUC) [16]. The ROC curve is a plot that illustrates the performance of a binary classifier as the discriminative threshold is varied. The curve is created by plotting the true positive rate against the false positive rate at various threshold settings.

Figure 4 presents a ROC graph. The ROC curves feature the true positive rate on the Y axis and the false positive rate on the X axis. This means that the top left corner of the plot is the 'ideal point' since this corresponds to a false positive rate of 0 and a true positive rate of 1. Larger AUC means better performance.

The reason why we use the ROC/AUC for evaluation is that, although it is necessary to increase the accuracy for the diagnosis of schizophrenia, it is of significance in reducing the false alarm rate because the false diagnosis is an extremely big mistake.

4.4. Experimental results

As shown in Table 1, we compare the ROC/AUC results of different systems. Since the SVM is used as classifier for all the systems, only the different models used for feature representation are listed in Table 1. Note that the baseline system denotes the trimming feature approach and that the result of the baseline system is obtained based on the code provided by the MLSP 2014 Classification Challenge. In addition, the setup of KCCA has a fixed gram matrix as introduced in [9].

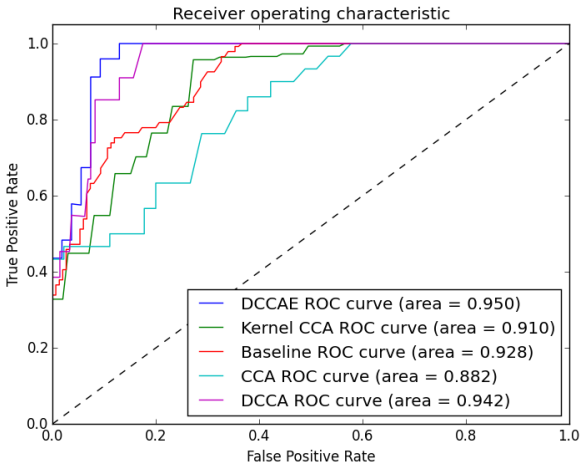

Figure 4: *ROC Curves for the Different Models.*

Figure 4 compares the different feature representation models for the diagnosis of schizophrenia and Table 1 shows the corresponding results. As shown in Table 1, DCCAE and DCCA models obtain much better results than the baseline system, but both KCCA and CCA cannot outperform the baseline system. These results mean that the bottleneck feature generated by DNN-based multi-view models can significantly improve the classification tasks for the diagnosis of schizophrenia, which is consistent with the results obtained in [7] in the sense that both the DCCA and the DCCAE derive a more robust multi-view feature representation than the KCCA and the CCA. Furthermore, the DCCAE obtains a better result than the DCCA, which suggests that the regularization based on auto-encoders for the deep CCA can further improve the representation of the bottleneck feature within the multi-modal features.

Category	ROC/AUC Results
DCCAE	**0.950**
DCCA	**0.942**
Baseline	0.928
KCCA	0.910
CCA	0.882

Table 1: Results obtained by the different feature representation models.

5. Conclusions

This work presents the deep multi-view models DCCA and DCCAE for the diagnosis of schizophrenia. Both the DCCA and the DCCAE combine deep neural networks with canonical correlation analysis such that the bottleneck features generated by the non-linear transformations of the DCCA and the DCCAE present maximal correlation of the two inputs. The experimental results suggest that the bottleneck features based on the DCCA and the DCCAE models outperform the trimming features used in the baseline system for the diagnosis of schizophrenia in terms of the ROC/AUC evaluation. In addition, the regularization based on auto-encoders for the deep CCA demonstrates that it delivers further improvement for the representation of the multi-modal features.

Future work will compare our methods with other public proposals for the diagnosis of schizophrenia.

6. References

[1] A. V. Lebedev, "The 10th annual MLSP competition: Second place," in *Proc. of MLSP 2014 Schizophrenia Classification Challenge*, 2014.

[2] J. van and S. Kapur, "Schizophrenia," *Lancet*, vol. 374, pp. 635–645, 2009.

[3] C. A. Ross, R. L. Margolis, S. A. Reading, M. Pletnikov, and J. T. Coyle, "Neurobiology of schizophrenia," *Neuron*, vol. 52, no. 1, pp. 139–153, 2006.

[4] M. S. Cetin, F. Christensen, C. C. Abbott, J. M. Stephen, A. R. Mayer, and J. M. Canive, "Thalamus and posterior temporal lobe show greater inter-network connectivity at rest and across sensory paradigms in schizophrenia," *Neuroimage*, vol. 97, pp. 117–126, 2014.

[5] J. M. Segall, E. A. Allen, R. E. Jung, E. B. Erhardt, S. K. Arja, and K. Kiehl, "Correspondence between structure and function in the human brain at rest," *Front Neuroinform*, vol. 6, pp. 6–10, 2012.

[6] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. of NIPS*, 2013, pp. 1247–1255.

[7] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. of ICML*, 2015, pp. 1083–1092.

[8] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, no. 3-4, pp. 321–377, 1936.

[9] R. Arora and K. Livescu, "Kernel CCA for multi-view learning of acoustic features using articulatory measurements," in *Proc. of Machine Learning in Speech and Language Processing*, 2012.

[10] S. Akaho, "A kernel method for canonical correlation analysis," in *Proc. of International Meeting of the Psychometric Society*, 2001.

[11] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[12] J. Ngiam, A. K. Kim, M. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. of ICML*, 2011, pp. 689–696.

[13] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. of NIPS*, 2012, pp. 1097–1105.

[15] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. of EMNLP*, 2013.

[16] J. A. Hanley, "The meaning and use of the area under a receiver operating characteristic curve," *Radiology*, vol. 143, no. 1, pp. 29–36.