Deep Learning-Based Feature Representation for AD/MCI Classification

Heung-Il Suk and Dinggang Shen

Department of Radiology and Biomedical Research Imaging Center (BRIC),
University of North Carolina at Chapel Hill
{hsuk,dgshen}@med.unc.edu

Abstract. In recent years, there has been a great interest in computer-aided diagnosis of Alzheimer's Disease (AD) and its prodromal stage, Mild Cognitive Impairment (MCI). Unlike the previous methods that consider simple low-level features such as gray matter tissue volumes from MRI, mean signal intensities from PET, in this paper, we propose a deep learning-based feature representation with a stacked auto-encoder. We believe that there exist latent complicated patterns, *e.g.*, non-linear relations, inherent in the low-level features. Combining latent information with the original low-level features helps build a robust model for AD/MCI classification with high diagnostic accuracy. Using the ADNI dataset, we conducted experiments showing that the proposed method is 95.9%, 85.0%, and 75.8% accurate for AD, MCI, and MCI-converter diagnosis, respectively.

1 Introduction

Alzheimer's Disease (AD), characterized by progressive impairment of cognitive and memory functions, and its prodromal stage, Mild Cognitive Impairment (MCI), are the most prevalent neurodegenerative brain diseases in the elderly subjects. A recent research by Alzheimer's Association reports that AD is the sixth-leading cause of death in the United States, rising significantly every year in terms of the proportion of cause of death [1]. Researchers in many scientific fields have devoted their efforts to understand the underlying mechanism that causes these diseases and to identify pathological biomarkers for diagnosis or prognosis of AD/MCI by analyzing different types of imaging modalities, such as Magnetic Resonance Imaging (MRI) [3], Positron Emission Tomography (PET) [11], functional MRI (fMRI) [5], etc.

Recent research has shown that it's beneficial to fuse complementary information from different modalities in discriminating AD/MCI patients from Healthy normal Controls (HC) [12]. For instance, Hinrichs *et al.* [6] and Zhang *et al.* [13], independently, utilized a kernel-based machine learning technique to combine the complementary information from multi-modal data. Furthermore, [13] proposed to select features by means of sparse representation, which jointly learn the tasks of clinical label identification and clinical scores prediction.

Although these researches presented the effectiveness of their methods in their own experiments on multi-modal AD/MCI classification, the main limitation of

K. Mori et al. (Eds.): MICCAI 2013, Part II, LNCS 8150, pp. 583–590, 2013.

the previous work is that they considered only simple low-level features such as gray matter tissue volumes from MRI, mean signal intensities from PET, and biological measures from CerebroSpinal Fluid (CSF). In this paper, we assume that there exists hidden or latent high-level information inherent in the original features, which can be helpful to build a more robust model.

For the past decade, a deep architecture [2] has gained a great attention in various fields due to its representational power. Motivated by the recent work [2, 8], we exploit deep learning for a feature representation, and ultimately to enhance classification accuracy. Specifically, a '*Stacked Auto-Encoder*' (SAE) is utilized to discover a latent representation from the neuroimaging and biological low-level features. To our best knowledge, this is the first work that considers deep learning for feature representation in brain disease diagnosis and prognosis. Our experimental results on ADNI dataset proves the effectiveness of the proposed method.

2 Materials and Preprocessing

In this work, we use the ADNI dataset publicly available on the web[1]. Specifically, we consider the baseline MRI, PET, and CSF data acquired from 51 AD patients, 99 MCI patients (43 MCI patients who progressed to AD, and 56 MCI patients who did not progress to AD in 18 months), and 52 healthy normal controls. Along with the brain image data, two types of clinical scores, Minimum Mental State Examination (MMSE) and Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog), are also provided for each subject.

The MRI and PET images were preprocessed by applying the typical procedures of anterior commissure-posterior commissure correction, skull-stripping, and cerebellum removal. We segmented MRI images into gray matter, white matter, and CSF, and then parcellated them into 93 Regions Of Interests (ROIs) based on Kabani *et al.*'s atlas [9]. The PET images were spatially normalized by coregistering them to their respective MRI images. For each ROI, we used the gray matter tissue volume from MRI and the mean intensity from PET as features, which are most widely used in the field for AD/MCI diagnosis [3,6,13]. Therefore, we have 93 features from a MRI image and the same dimensional features from a PET image. In addition, we have 3 CSF biomarkers of Aβ_{42}, t-tau, and p-tau.

3 Methods

Fig. 1 illustrates a schematic diagram of the proposed method. Given multimodal data along with the class-label and clinical scores, we first extract features from MRI and PET as explained in Section 2. We then discover a latent feature representation from the low-level features in MRI, PET, and CSF, independently, by deep learning with SAE. A multi-task learning on the augmented feature vectors, *i.e.*, concatenation of the original low-level features and the SAE-learned

[1] Available at `http://www.loni.ucla.edu/ADNI`

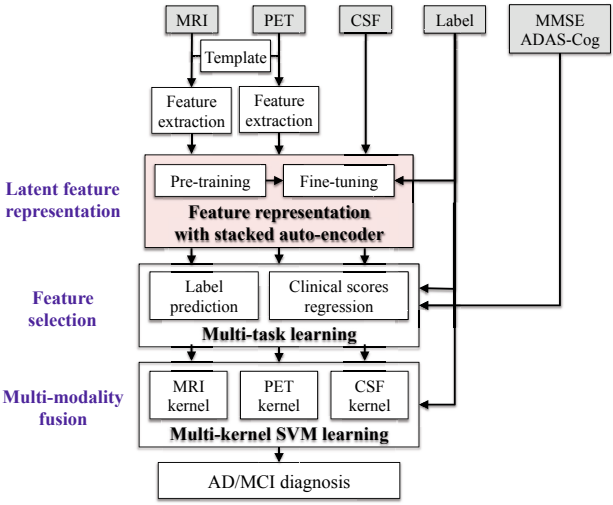


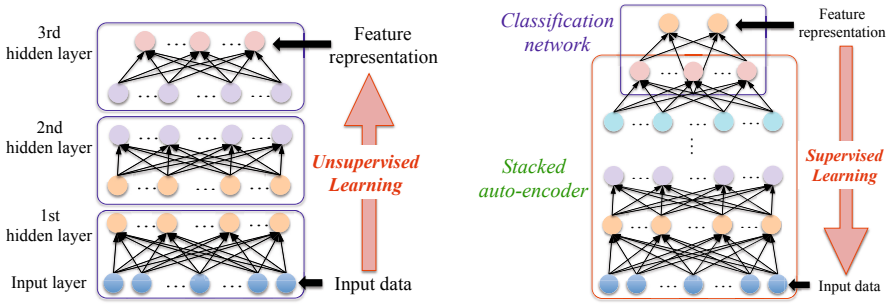
Fig. 1. An illustration of the proposed method for AD/MCI diagnosis

features, is applied to select features that jointly represent the class label and the clinical scores. Finally, we fuse the selected multi-modal feature information with a multi-kernel Support Vector Machine (SVM).

3.1 Stacked Auto-encoder

Auto-encoder is one type of artificial neural networks structurally defined by three layers: input layer, hidden layer, and output layer. The aim of the auto-encoder is to learn a latent or compressed representation of the input vector $\mathbf{x}$. Let D_H and D_I denote, respectively, the number of hidden and input units. Given an input vector $\mathbf{x} \in \mathbb{R}^{D_I}$, an auto-encoder maps it to a latent representation $\mathbf{y}$ through a deterministic mapping $\mathbf{y} = f(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$, parameterized by the weight matrix $\mathbf{W}_1 \in \mathbb{R}^{D_H \times D_I}$ and the bias vector $\mathbf{b}_1 \in \mathbb{R}^{D_H}$. The representation $\mathbf{y} \in \mathbb{R}^{D_H}$ from the hidden layer is then mapped back to a vector $\mathbf{z} \in \mathbb{R}^{D_I}$, which approximately reconstructs the input vector $\mathbf{x}$ by another deterministic mapping $\mathbf{z} = \mathbf{W}_2\mathbf{y} + \mathbf{b}_2 \approx \mathbf{x}$, where $\mathbf{W}_2 \in \mathbb{R}^{D_I \times D_H}$ and $\mathbf{b}_2 \in \mathbb{R}^{D_I}$. In this study, we consider a logistic sigmoid function for $f(a) = 1/\left(1 + exp(-a)\right)$.

Recent studies in machine learning have shown that a deep or hierarchical architecture is useful to find highly non-linear and complex patterns in data [2]. Motivated by the studies, in this paper, we consider SAE, in which an auto-encoder becomes a building block, for a feature representation in neuroimaging or biological data. Thanks to its hierarchical nature in structure, one of the most important characteristics of the SAE is to learn or discover patterns such as non-linear relations among input values. Utilizing its representational power, we find a latent representation of the original low-level features extracted from neuroimaging or biological data. Note that in order to obtain highly non-linear



(a) Unsupervised pre-training (b) Supervised fine-tuning

Fig. 2. A deep architecture of our stacked auto-encoder and the two-step parameter optimization scheme

relations, we allow the hidden layers to have any number of units, even larger than the input dimension, from which we can still find an interesting structure by imposing a sparsity constraint on the hidden units, which is called a sparse auto-encoder [10]. Specifically, we penalize a large average activation of a hidden unit over the training samples. This penalization drives many of the hidden units' activation to be zero, resulting in sparse connections between layers.

With regard to training a SAE hierarchical network, the conventional gradient-based optimization starting from random initialization suffers from falling into a poor local optimum. Recently, Hinton *et al.*, introduced a greedy layer-wise unsupervised learning algorithm and showed its success to learn a deep belief network [7]. The key concept in a greedy layer-wise learning is to train one layer at a time. That is, we first train the 1^{st} hidden layer with the training data as input, and then train the 2^{nd} hidden layer with the outputs from the 1^{st} hidden layer as input, and so on. That is, the representation of the l-th hidden layer is used as input for the $(l+1)$-th hidden layer. This greedy layer-wise learning is called '*pre-training*' (Fig. 2(a)). It is worth noting that the pre-training is performed in an unsupervised manner.

To improve diagnostic performance in AD/MCI identification, we further optimize the deep network in a supervised manner. Accordingly, we stack another output layer on top of the SAE. This top output layer is used to represent the class label of an input data. We set the number of units in the output layer to be equal to the number of classes of interest. This extended network can be considered as a traditional multi-layer neural network and, in this paper, we call it SAE-classifier. Therefore, it is straightforward to optimize the deep network by back-propagation with gradient descent, having parameters, except for the last classification network, initialized by the pre-trained ones. This supervised optimization step is called '*fine-tuning*' (Fig. 2(b)). From an optimization point of view, it is known that the parameters obtained from the pre-training step helps the fine-tuning optimization to reduce the risk of falling into a poor local optimum [7]. This makes the deep learning distinguished from the conventional multi-layer neural network.

Besides the fine-tuning, we also utilize the top output layer to determine the optimal SAE structure, for which the solution is combinatorial. In this paper, we apply a grid search and choose a network structure that produces the best classification accuracy. Once we determine the SAE structure, we consider the outputs from the last hidden layer as our latent feature representation. By concatenating the SAE-learned feature representation with the original low-level features, we construct an augmented feature vector that is then fed into the multi-task learning as explained below.

3.2 Multi-task and Multi-kernel SVM Learning

Following Zhang and Shen's work [13], we consider the multi-task learning for feature selection. Let $m \in \{1, \cdots, M\}$ denote a modality index, $s \in \{1, \cdots, S\}$ denote a task index[2], $\mathbf{t}_s^{(m)}$ denote a target response vector, and $\mathbf{F}^{(m)} \in \mathbb{R}^{N \times D}$ denote a set of the augmented feature vectors, where N and D are, respectively, the number of samples and the dimension of the augmented feature vectors. In the multi-task learning, we focus on finding optimal weight coefficients $\mathbf{a}_s^{(m)}$ to regress the target response vector with a combination of the features in $\mathbf{F}^{(m)}$ with a group sparsity constraint as follows:

$$J\left(\mathbf{A}^{(m)}\right) = \min_{\mathbf{A}^{(m)}} \frac{1}{2} \sum_{s=1}^{S} \left\| \mathbf{t}_s^{(m)} - \mathbf{F}^{(m)} \mathbf{a}_s^{(m)} \right\|_2^2 + \lambda \left\| \mathbf{A}^{(m)} \right\|_{2,1} \tag{1}$$

where $\mathbf{A}^{(m)} = \left[\mathbf{a}_1^{(m)} \cdots \mathbf{a}_s^{(m)} \cdots \mathbf{a}_S^{(m)} \right]$ and λ is a sparsity control parameter. In Eq. (1), $\left\| \mathbf{A}^{(m)} \right\|_{2,1} = \sum_{d=1}^{D} \left\| \mathbf{A}^{(m)}[d] \right\|_2$, where $\mathbf{A}^{(m)}[d]$ denotes the d-th row of the matrix $\mathbf{A}^{(m)}$. This $l_{2,1}$-norm imposes to select features that are jointly used to regress the target response vector $\{\mathbf{t}_s^{(m)}\}_{s=1}^{S}$ across tasks[3]. We select features whose absolute weight coefficient is larger than zero for SVM learning.

Given the feature-selected training samples $\tilde{\mathbf{X}}^{(m)} = \{\tilde{\mathbf{x}}_i^{(m)}\}_{i=1}^{N}$ and the test sample of $\tilde{\mathbf{x}}^{(m)}$ from modalities $m \in \{1, \cdots, M\}$, the decision function of the multi-kernel SVM is defined as follows:

$$f\left(\tilde{\mathbf{x}}^{(1)}, \cdots, \tilde{\mathbf{x}}^{(M)}\right) = sign\left\{ \sum_{i=1}^{N} \zeta_i \alpha_i \sum_{m=1}^{M} \beta_m k^{(m)}\left(\tilde{\mathbf{x}}_i^{(m)}, \tilde{\mathbf{x}}^{(m)}\right) + b \right\} \tag{2}$$

where ζ_i is the class-label of the i-th sample, α_i and b are, respectively, a Lagrangian multiplier and a bias, $k^{(m)}\left(\tilde{\mathbf{x}}_i^{(m)}, \tilde{\mathbf{x}}^{(m)}\right) = \phi^{(m)}\left(\tilde{\mathbf{x}}_i^{(m)}\right)^T \phi^{(m)}\left(\tilde{\mathbf{x}}^{(m)}\right)$ is a kernel function of the m-th modality, $\phi^{(m)}$ is a kernel-induced mapping function, and $\beta_m \geq 0$ is a weight coefficient of the m-th modality with the constraint of $\sum_m \beta_m = 1$. Refer to [4, 13] for a detailed explanation.

[2] In our case, the tasks are to predict class-label, MMSE, and ADAS-Cog scores.

[3] In this work, $\mathbf{t}_s^{(1)} = \cdots = \mathbf{t}_s^{(m)} = \cdots = \mathbf{t}_s^{(M)}$.

4 Experimental Results and Discussions

We consider three binary classification problems: AD vs. HC, MCI vs. HC, and MCI Converter (MCI-C) vs. MCI Non-Converter (MCI-NC). In the experiment of the MCI vs. HC classification, both MCI-C and MCI-NC data were used for the MCI class. For each classification problem, we applied a 10-fold cross validation. Specifically, we randomly partitioned the dataset into 10 subsets and then used 9 out of 10 subsets for training and the remaining one for test. In order to determine the hyper-parameters of λ in Eq. (1) and β in Eq. (2), another round of cross-validation was performed within the training data. We repeated these whole process 10 times for unbiased evaluation. We used a linear kernel in SVM.

In order to show the validity of the SAE-learned Feature representation (SAEF), we compared the results of the proposed method with those from the original Low-Level Features (LLF) using the same strategies of feature selection and classifier learning. We should note that, for fair comparison, we used the same training and test data across the experiments for all the competing methods.

With regard to the SAE structure, we considered three hidden layers for MRI, PET, and CONCAT, and two hidden layers for CSF, which were determined based on our preliminary experiments. Here, CONCAT represents the concatenation of the MRI, PET, and CSF features into a single vector. As explained in Section 3.1, we determined the number of hidden units based on the classification results with a SAE-classifier. The classification accuracies and the optimal structure of the SAE-classifier are shown in Table 1. We used a DeepLearnToolbox[4] to train the SAE, and a SLEP toolbox[5] for the multi-task learning, respectively.

Table 2 presents the mean classification accuracies of the competing methods. The method of multi-kernel SVM with LLF corresponds to Zhang and Shen's method [13]. Although the approach based on the augmented feature vector (LLF+SAEF) with a single-modality was outperformed for some cases by the LLF-based one, the proposed method with a Multi-Kernel SVM (MK-SVM) produced the best performances for AD vs. HC, MCI vs. HC, and MCI-C vs. MCI-NC classification problems, with the accuracies of 95.9%, 85.0%, and 75.8%, respectively. It should be noted that the performance improvement by the proposed method was 4.0% for MCI-C vs. MCI-NC classification, which is the most important for early diagnosis and treatment.

In order to further validate the effectiveness of the proposed method, we also computed a statistical significance of the results with paired t-test: AD vs. HC (0.0127), MCI vs. HC (0.0568), and MCI-C vs. MCI-NC (0.0096). The test was performed with the results obtained from Zhang and Shen's method [13] (LLF with MK-SVM) and the proposed method (LLF+SAEF with MK-SVM). The proposed method statistically outperformed Zhang and Shen's method, especially for AD vs. HC (0.0127) and MCI-C vs. MCI-NC (0.0096).

[4] Available at '`https://github.com/rasmusbergpalm/DeepLearnToolbox`'
[5] Available at '`http://www.public.asu.edu/~jye02/Software/SLEP/index.htm`'

Table 1. Performance of the SAE-classifier (mean±standard deviation). '# units' denotes the number of hidden units (bottom-to-top layer) that produced the corresponding performance.

		MRI	PET	CSF	CONCAT
AD vs. HC	Accuracy	0.857±0.018	0.859±0.021	0.831±0.016	0.899±0.014
	# units	500-50-10	1000-50-30	50-3	500-100-20
MCI vs. HC	Accuracy	0.706±0.021	0.670±0.018	0.683±0.020	0.737±0.025
	# units	100-100-20	300-50-10	10-3	100-50-20
MCI-C vs. MCI-NC	Accuracy	0.549±0.037	0.595±0.044	0.589±0.026	0.602±0.031
	# units	100-100-10	100-100-10	30-2	500-50-20

Table 2. Performance comparison of the competing methods. The method of LLF with MK-SVM corresponds to Zhang and Shen's work [13]. (SK: Single-Kernel, MK: Multi-Kernel).

			Features		
			LLF	SAEF	LLF+SAEF
AD vs. HC	SK-SVM	MRI	0.817±0.018	0.802±0.033	0.823±0.025
		PET	0.821±0.017	0.834±0.016	0.838±0.021
		CSF	0.720±0.017	0.763±0.055	0.799±0.015
		CONCAT	0.893±0.019	0.832±0.027	0.853±0.032
	MK-SVM		0.945±0.008	0.939±0.018	**0.959±0.011**
MCI vs. HC	SK-SVM	MRI	0.732±0.018	0.673±0.015	0.740±0.021
		PET	0.702±0.032	0.673±0.031	0.682±0.033
		CSF	0.640±0.021	0.660±0.020	0.680±0.012
		CONCAT	0.737±0.017	0.701±0.028	0.769±0.023
	MK-SVM		0.840±0.011	0.792±0.024	**0.850±0.012**
MCI-C vs. MCI-NC	SK-SVM	MRI	0.568±0.026	0.542±0.034	0.550±0.027
		PET	0.626±0.036	0.606±0.034	0.592±0.034
		CSF	0.527±0.026	0.581±0.029	0.574±0.015
		CONCAT	0.616±0.043	0.584±0.041	0.603±0.023
	MK-SVM		0.718±0.026	0.735±0.024	**0.758±0.020**

We should mention that the data fusion in deep learning was considered through the concatenation of the features from multiple modalities. But, it is limited as a shallow model to discover the non-linear relations among modalities. We believe that although the proposed SAE-based method is successful to find latent information, resulting in performance enhancement, there is still a room to design a multi-modal deep network for the shared representation among modailities. It is also an important issue how to efficiently interpret or visualize the trained weights of the deep network in brain research.

5 Conclusion

We propose a deep learning-based feature representation for AD/MCI diagnosis. Unlike the previous methods that consider only simple low-level features

extracted directly from neuroimages, the proposed method can successfully discover latent feature representation such as non-linear correlations among features that improve diagnosis accuracy. Using the ADNI dataset, we evaluated the performance of the proposed method and compared against the state-of-the-art method [13]. The proposed method outperformed the competing method and presented the accuracies of 95.9%, 85.0%, and 75.8% for AD, MCI, and MCI-C diagnosis, respectively.

References

1. Alzheimer's Association: 2012 Alzheimer's disease facts and figures. Alzheimer's & Dementia 8(2), 131–168 (2012)
2. Bengio, Y.: Learning deep architectures for AI. Foundations and Trends in Machine Learning 2(1), 1–127 (2009)
3. Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q.: Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. Neurobiology of Aging 32(12), 2322.e19–2322.e27 (2011)
4. Gönen, M., Alpaydin, E.: Multiple kernel learning algorithms. Journal of Machine Learning Research 12, 2211–2268 (2011)
5. Greicius, M.D., Srivastava, G., Reiss, A.L., Menon, V.: Default-mode network activity distinguishes Alzheimer's disease from healthy aging: Evidence from functional MRI. PNAS 101(13), 4637–4642 (2004)
6. Hinrichs, C., Singh, V., Xu, G., Johnson, S.C.: Predictive markers for AD in a multimodality framework: An analysis of MCI progression in the ADNI population. NeuroImage 55(2), 574–589 (2011)
7. Hinton, G.E., Osindero, S., Teh, Y.W.: A fast learning algorithm for deep belief nets. Neural Computation 18(7), 1527–1554 (2006)
8. Hinton, G.E., Salakhutdinov, R.R.: Reducing the dimensionality of data with neural networks. Science 313(5786), 504–507 (2006)
9. Kabani, N., MacDonald, D., Holmes, C., Evans, A.: A 3D atlas of the human brain. NeuroImage 7(4), S717 (1998)
10. Larochelle, H., Bengio, Y., Louradour, J., Lamblin, P.: Exploring strategies for training deep neural networks. Journal of Machine Learning Research 10, 1–40 (2009)
11. Nordberg, A., Rinne, J.O., Kadir, A., Langstrom, B.: The use of PET in Alzheimer disease. Nature Reviews Neurology 6(2), 78–87 (2010)
12. Perrin, R.J., Fagan, A.M., Holtzman, D.M.: Multimodal techniques for diagnosis and prognosis of Alzheimer's disease. Nature 461, 916–922 (2009)
13. Zhang, D., Shen, D.: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59(2), 895–907 (2012)