DOI: 10.1002/alz.092330

BIOMARKERS

POSTER PRESENTATION

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

NEUROIMAGING

'Diet' deep canonical correlation analysis for high-dimensional genetics study of brain imaging phenotypes in Alzheimer's disease

Shu Yang[1] | Austin Wang[1] | Jingxuan Bao[1] | Shizhuo Mu[1] | Yanbo Feng[1] | Zixuan Wen[1] | Jae Young Baik[1] | Junhao Wen[1] | Bojian Hou[1] | Rongguang Wang[1] | Heng Huang[2] | Andrew J. Saykin[3] | Paul M. Thompson[4] | Christos Davatzikos[1] | Li Shen[1] | for the ADNI and AI4AD

[1]University of Pennsylvania, Philadelphia, PA, USA

[2]University of Maryland, College Park, MD, USA

[3]Indiana University, Indianapolis, IN, USA

[4]University of Southern California, Los Angeles, CA, USA

Correspondence
Shu Yang, University of Pennsylvania, Philadelphia, PA, USA.
Email: shu.yang@pennmedicine.upenn.edu

Abstract

Background: Understanding the relationship between genetic variations and brain imaging phenotypes is an important issue in Alzheimer's disease (AD) research. As an alternative to GWAS univariate analyses, canonical correlation analysis (CCA) and its deep learning extension (DCCA) are widely used to identify associations between multiple genetic variants such as SNPs and multiple imaging traits such as brain ROIs from PET/MRI. However, with the recent availability of numerous genetic variants from genotyping and whole genome sequencing data for AD, these approaches often suffer from severe overfitting when dealing with 'fat' genetics data, e.g. large numbers of SNPs with much smaller numbers of samples.

Methods: Here, we propose to tackle the challenge by integrating an efficient model parameterization approach from Mila's Diet Network architecture into DCCA to handle high dimensional SNP data in AD imaging-genetics study (Figure 1). The new method, DietDCCA, was applied to nine datasets derived from 955 subjects in the ADNI data. Each dataset contains 68 FreeSurfer cortical ROIs from the florbetapir (AV45) PET imaging and varied numbers of SNPs from 810 to 11,938 based on different significance thresholds derived from previous studies.

Results: Firstly, we compared our DietDCCA with DCCA on each of the nine datasets to demonstrate the improvement on test correlations (Figure 2). DietDCCA outperformed DCCA by a large margin on all datasets even when the SNPs are as many as ~10k. Next, we sought to verify if the detected correlations were contributed by meaningful SNPs and extracted the SNP feature that has the largest weight in each neuron of the diet net layer (Figure 3). DietDCCA successfully selected the APOE4 SNP

(rs429358) for most cases and also picked out SNPs in other genes (ABCA7, APOC2, CLPTM1, NECTIN2) that were previously reported to associate with AD.

Conclusions: We introduced a novel method, DietDCCA, to handle high-dimensional SNP features in AD imaging-genetics study. The initial investigation of DietDCCA on the ADNI data showed promises in detecting correlation signals with AV45 ROIs from biologically meaningful SNPs. The study supplies a novel and effective tool to study the genetic basis of AD imaging phenotypes for future analyses.

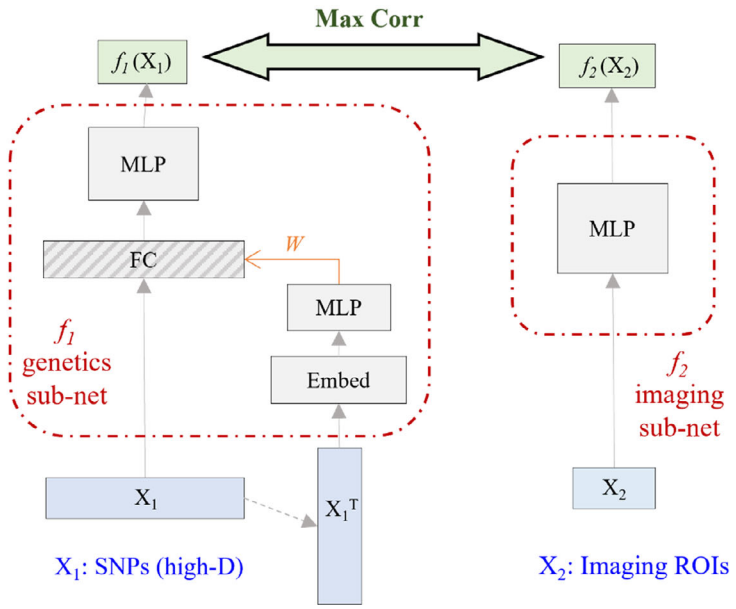

Fig. 1: Schematic overview of the diet deep CCA (DietDCCA) model. The proposed model consists of two sub-neural networks, one with 'diet' net for the high-dimensional genetics data modality and one standard multilayer perceptron (MLP) for the imaging modality, learned jointly so that the outputs from the two sub-nets are maximally correlated. For the diet net component in the genetics sub-net, a small neural network was learned to transform a distributed representation of each SNP genotype feature to the vector of weights for that feature in the big neural network of the original deep CCA.

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

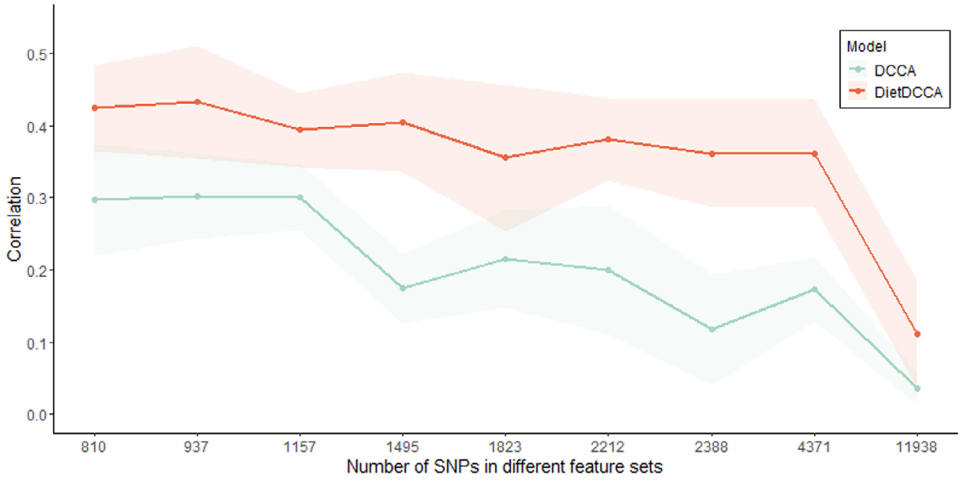

Fig. 2: Performance of the proposed DietDCCA compared to the deep CCA (DCCA) model. Each dot represents the test correlation on one of the 9 datasets with varied numbers of SNP features from 810 to 11,938. The correlation values are averaged across five different training-test splits to avoid bias, with the std shown as the shades in the plot. Both models were fine-tuned with classic regularization techniques (weight decay, dropout, early stopping, etc.) on the validation set.

Top1 SNP per neuron	Count	Associated gene (dbSNP)
rs429358	90	APOE4
rs7259679	7	APOC4-APOC2
rs3764642	6	CLPTM1
rs384973	3	ABCA7
rs8111069	3	NECTIN2
rs144261139	3	CLPTM1
rs117316645	2	NECTIN2
rs403729	1	CLPTM1
rs440277	1	NECTIN2
rs4803781	1	NECTIN2
rs59136988	1	CLPTM1
rs8100120	1	CLPTM1
rs7259679	1	CLPTM1

Fig. 3: The top1 important features picked out by each neuron in the final DietDCCA model for the 810 SNPs dataset. There were, in total, 120 neurons in the diet net layer of DietDCCA. 90 of them attributed the highest weights to the APOE4 SNP feature while the rest of the neurons also picked out SNPs in some other AD related genes. The SNPs-genes information was obtained from dbSNP.