Multi-Modality Disease Modeling via Collective Deep Matrix Factorization

Qi Wang
Michigan State University
wangqi19@msu.edu

Mengying Sun
Michigan State University
sunmeng2@msu.edu

Liang Zhan
University of Wisconsin-Stout
zhanl@uwstout.edu

Paul Thompson
University of Southern California
pthomp@usc.edu

Shuiwang Ji
Washington State University
sji@eecs.wsu.edu

Jiayu Zhou
Michigan State University
jiayuz@msu.edu

ABSTRACT

Alzheimer's disease (AD), one of the most common causes of dementia, is a severe irreversible neurodegenerative disease that results in loss of mental functions. The transitional stage between the expected cognitive decline of normal aging and AD, mild cognitive impairment (MCI), has been widely regarded as a suitable time for possible therapeutic intervention. The challenging task of MCI detection is therefore of great clinical importance, where the key is to effectively fuse predictive information from multiple heterogeneous data sources collected from the patients. In this paper, we propose a framework to fuse multiple data modalities for predictive modeling using deep matrix factorization, which explores the non-linear interactions among the modalities and exploits such interactions to transfer knowledge and enable high performance prediction. Specifically, the proposed collective deep matrix factorization decomposes all modalities simultaneously to capture non-linear structures of the modalities in a supervised manner, and learns a modality specific component for each modality and a modality invariant component across all modalities. The modality invariant component serves as a compact feature representation of patients that has high predictive power. The modality specific components provide an effective means to explore imaging genetics, yielding insights into how imaging and genotype interact with each other non-linearly in the AD pathology. Extensive empirical studies using various data modalities provided by Alzheimer's Disease Neuroimaging Initiative (ADNI) demonstrate the effectiveness of the proposed method for fusing heterogeneous modalities.

CCS CONCEPTS

•Information systems →Data mining; •Computing methodologies →*Machine learning;*

KEYWORDS

Mild cognitive impairment, multi-modality, deep neural network, matrix factorization, medical imaging, genetics.

KDD'17, August 13–17, 2017, Halifax, NS, Canada.

DOI: http://dx.doi.org/10.1145/3097983.3098164

ACM Reference format:
Qi Wang, Mengying Sun, Liang Zhan, Paul Thompson, Shuiwang Ji, and Jiayu Zhou. 2017. Multi-Modality Disease Modeling via Collective Deep Matrix Factorization. In *Proceedings of KDD'17, August 13–17, 2017, Halifax, NS, Canada.*, 10 pages.
DOI: http://dx.doi.org/10.1145/3097983.3098164

1 INTRODUCTION

Alzheimer's disease (AD) is a severe neurodegenerative disease causing 60% to 70% dementia [43]. It starts with vanished memory and progresses to an advanced stage followed by cognitive function loss, which ultimately leads to death. Currently, AD ranks the sixth leading cause of death in the U.S. and the number of patients affected is expected to reach 13.4 million by the year 2050, which induces substantial burden on the healthcare system [4]. The transitional stage between expected cognitive decline of normal aging and AD, mild cognitive impairment (MCI) has been considered as suitable for possible early therapeutic intervention for AD [34]. However, it is not yet possible to determine the underlying cause of MCI from symptoms of a person [5]. Effective prognosis of MCI can greatly benefit public health and reduce healthcare burden.

Variants in multiple biological measures such as medical imaging and genotype can provide complementary information on brain structure and function, thus improve capability in differentiating between normal aging subjects and MCI patients. Magnetic resonance imaging (MRI), providing detailed anatomical description of the brain, has been extensively used in extracting imaging biomarkers and identifying MCI subjects [19, 30, 39, 40]. T1-weighted MRI (T1 MRI) can capture structural information of gray matter in the brain, while diffusion-weighted MRI (dMRI) is sensitive to microscopic properties of brain's white matter. Combining T1 MRI and dMRI together provides a comprehensive illustration of the brain than utilizing them separately. Moreover, prior studies strongly favor a joint analysis on multiple modalities including imaging and genetics, since it has been shown that genetic variants have played a significant role in the onset of the disease [7, 36, 44, 45].

However, few prior studies combined two types of MRI imaging in detecting MCI, let alone a joint model that incorporate imaging modalities and genetic information. One significant challenge is the limited sample size. It is usually very costly to construct large cohort studies that involve imaging and genetic data. For example, more than $60 million has been devoted to the first stage of Alzheimer's Disease Neuroimaging Initiative (ADNI) to collect 819 subjects' brain imaging data, genetic data and other biological

samples. Different biological data modalities have different feature dimensions. For example, imaging data contains hundreds to thousands features, while the feature dimension of genetic data is around 1 million. Due to the high dimensionality of brain images and genetic markers, directly combining multiple modalities will increase the feature dimension drastically, which not only makes it difficult to extract valid predictive signals, but also induces overfitting problems. Also, some subjects do not have genetic data or dMRI data because they did not participate some parts of the study. Directly combining multiple modalities means those subjects must be discarded, which significantly reduces the sample size. Moreover, different modalities describe different aspects of brain: T1 MRI captures areas composed of neurons while dMRI estimates connection between those areas; the genotype impacts the disease in a way that is not directly related to brain structure and function. As such, all these data modalities are interacting in a complicated manner, suggesting that directly combining feature spaces may not lead to effective integration.

Analysis of high dimensional data can greatly benefit from its intrinsic low-rank structures, exploiting which allows us to significantly reduce the feature dimensionality while maintain most information in data. Recent studies have identified such low-rank properties in imaging and genetic data [28, 42, 49]. Matrix factorization techniques [23, 25] are powerful tools to recover the low rank structure of a matrix and have been widely used in many data mining and machine learning applications. Because of its capability to denoise data, such approach is especially attractive in processing noisy data such as genetics and imaging. Matrix factorization also provides an integrated approach to fuse multiple data modalities by mapping different modalities to a shared subspace. This method has been widely applied in network analysis [9] and clustering [3]. We note that matrix factorization techniques have a strong linear assumption that objects interact with each other linearly in a low dimensional subspace. However, brain as well as genotype-phenotype interactions have inherent complex structure [15, 17, 24]. For example, it has been identified that human brain functional networks have a hierarchical modular organization structure [29]. Thus, the linear assumption in traditional matrix factorization may fail to capture the complexity, nonlinearities and hierarchical interactions among different modalities in AD research.

In this paper, we propose a deep matrix factorization framework to fuse information from multiple modalities and transfer predictive knowledge in order to differentiate MCI patients from cognitive normal subjects. Specifically, we build a nonlinear hierarchical deep matrix factorization framework which decomposes each modality into a *modality invariant component* and a *modality specific component* guided by supervision information. The proposed collective deep matrix factorization delivers higher predictive performance than its linear counterpart, since its deep nonlinear structure can discover the hidden complexity and nonlinearity of original data, and map original data which are not linear separable into a representation that can make subjects easier to be separated. Moreover, the modality specific term can be used to uncover complicated interactions among different modalities that cannot be discovered by traditional matrix factorization methods. We perform extensive empirical studies on ADNI dataset to identify MCI patients by fusing three modalities including T1 MRI, dMRI, and genotype. We also compare our method with state-of-the-art deep multimodality algorithms including deep neural network, DCCA [6] and DCCAE [41]. The results demonstrate the effectiveness of the proposed approach.

2 RELATED WORK

The proposed framework performs collective deep matrix factorization for multi-modal learning. In this section, we review the related work in fields of multi-modal learning and matrix factorization, and point out the differences.

2.1 Learning from multiple modalities

Fusing information from multiple modalities has been an active research topic that attracts intensive efforts in the community. Deep learning techniques have recently been utilized to fuse multiple data modalities. In [32], the authors proposed multi-modal deep learning. They extracted shared representations of two modalities by reconstructing them from the one modality that was available at testing and showed their method can learn descent cross modality features. In [6], traditional canonical correlation analysis was extended to deep canonical correlation analysis by learning two deep neural networks on two modalities and maximizing the canonical correlation of the output of two neural networks. The authors in [41] combined the ideas from [32] and [6], and proposed deep canonically correlated auto-encoders that optimize the combination of canonical correlation between the learned representation of two modalities and the reconstruction errors of auto-encoders.

Multi-modal learning is closely related to multi-view learning. Multi-view learning can be considered as a special setting of multi-modal where the multiple modalities have the same set of samples. [16] introduced a multi-view deep learning model in recommendation systems which mapped users and items into a shared semantic space by deep neural network and recommended items which have maximum similarity with users in the mapped space. In [20], the authors presented a novel multi-view deep network to deal with multi-view data. The method learned a discriminant and view-invariant representation shared between multiple views using a non-linear deep neural network.

All these methods try to remove the modality specific correlations layer-by-layer until a modality invariant representation is learned. Our method also tries to learn a modality invariant representation. However, instead of directly removing all the modality specific information, we decompose the input modality-full data into a modality invariant term and several modality specific terms. This process shows how modality specific information is separated from modality invariant part explicitly. Also, the modality specific terms can be used to analyze the relationship between multiple modalities and provide their associations.

2.2 Matrix factorization

Matrix factorization techniques exploit the inherent low-rank structure of a matrix and learn latent representations of objects involved. Recent years have witnessed growing efforts in improving matrix

factorization using deep learning techniques. In [38], authors extended semi-non-negative matrix factorization to deep semi-non-negative matrix factorization to perform clustering. Deep semi-non-negative matrix factorization can learn different attributes of the data from different hidden layers and cluster data by those different attributes. In [37], the authors proposed multi-layer non-negative matrix factorization network for classification task. They stacked non-negative matrix factorization into several layers and took step-by-step approach in learning the features, which can provide intuitive explanation of learning steps in each layer. [35] presented a method based on DNN and applied low-rank matrix factorization on the final weight layer. Low-rank matrix factorization reduced the number of parameters involved in the network. Thus, it reduced training time without a significant loss in final recognition accuracy compared to a full-rank representation.

Matrix factorization has also been used to fuse multiple data modalities. [9] proposed factorized similarity learning to mining the similarity between pairs of nodes in a network from multiple modalities. They fused knowledge from network structure (links), content, and user supervision to achieve stable and generalized similarity learning on networks by matrix factorization.

Although previous work has been worked on extracting high-level representation or fusing multiple modalities by matrix factorization, none of them can combine the two methods to extract complex non-linear structures of multiple modalities and fuse learned high-level representations to improve performance.

3 FUSING MULTIPLE MODALITIES VIA DEEP MATRIX FACTORIZATION

3.1 Matrix factorization

Classical matrix factorization seeks to approximate a matrix with a low-rank matrix, by explicitly learning the matrix factors. Given a data matrix $X \in \mathbb{R}^{m\times n}$, matrix factorization learns two reduced matrix factors $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$, such that $X \approx UV^T$, and $r < \min(m, n)$ is the upper bound of the rank of the approximated matrix UV^T (the rank of UV^T can be less than r if columns of U or V are linearly dependent). The factors U and V are typically learned via an objective function:

$$\min_{U,V} \; d(X, UV^T), \text{ s.t. } U \in \mathcal{S}_1, V \in \mathcal{S}_2, \tag{1}$$

where $d(X, Y)$ is a distance metric function measuring the difference between matrices X and Y, and S_1 and S_2 are two constrains imposed on the factor matrices X and Y.

Typically the distance metric $d(X, Y)$ is chosen to be the Frobenius norm of the difference between X and Y. However, when missing values present in X, $d(X, Y)$ can be defined as the squared ℓ_2 distance between all the observed elements in X and their corresponding elements in Y. As such, we are able to learn matrix factors even with missing values, and the learned matrix factors can then be used to estimate the missing values under the low-rank assumption. This is the setup for matrix completion [8] and is commonly used in recommender systems [23]. The constraints $\mathcal{S}_1$ and $\mathcal{S}_2$ specify the feasible regions of the matrix factors to induce many desired properties, such as non-negativity $\mathcal{S} = \{U|U_{i,j} \geq 0, \forall i, j\}$ in non-negative matrix factorization [26] and sparsity $\mathcal{S} = \{U|\|U\|_1 \leq z\}$ for interpretable factors [48]. In addition, the complexity control can be implemented using Frobenius constraints $\mathcal{S} = \{U|\|U\|_F^2 \leq z\}$, which are equivalent to the Frobenius norm regularizations [22].

3.2 Collective matrix factorization for multi-modal analysis

The approximation in (1) addresses important semantics in data analysis. When the data matrix X describes the relationship between two types of entities, the factors U and V can be thought of as latent features or latent representations of the entities. For example, in recommender systems we use $X_{i,j}$ to describe the relationship (e.g., rating) between a user i and an item j. The row vector $\mathbf{u}^i \in \mathbb{R}^r$ gives a r-dimensional latent feature representation for the user i and similarly the row vector $\mathbf{v}^j \in \mathbb{R}^r$ is a latent representation of the item j. The two types of latent profile interact with each other linearly in the latent subspace $\mathbb{R}^r$, i.e., the observed relationship in $X_{i,j}$ can be explained as $\mathbf{u}^i(\mathbf{v}^j)^T$.

The latent representation/subspace perspective of matrix factorization allows us to link multiple data modalities, when the entities involved in the modalities are overlapped. In multi-modality modeling, assume we have two datasets $X_1 \in \mathbb{R}^{n\times d_1}$ and $X_2 \in \mathbb{R}^{n\times d_2}$ describing the same set of objects from two sets of features. For example, we study a set of n patients. X_1 includes d_1 features from T1 MRI modality and X_2 includes d_2 features from dMRI modality. Then we can apply the matrix factorization procedure to factorize both datasets and connect the two factorizations by enforcing a shared patient latent representation:

$$\min_{U,V_1,V_2} \; d(X_1, UV_1^T) + d(X_2, UV_2^T), \text{ s.t. } U \in \mathcal{S}_0, V_i \in \mathcal{S}_i, i = 1, 2,$$

where the latent representation U is thus jointly learned from two modalities. We call this U matrix *modality invariant*, as the representation now captures intrinsic properties of the patients. When performing regression and classification on patients, instead of using features from raw data matrices X_1 and X_2, we can use the latent representation. We can easily generalize this approach to handle more data modalities.

3.3 Capturing complex interactions via collective deep matrix factorization

One essential assumption associated to the classical matrix factorization is the linear dependence in the matrix. Therefore, it implicitly specifies that the latent representations learned from collective matrix factorization have to interact with each other linearly in the learned latent subspace. However, this assumption is too restrictive in many applications, especially in the modeling of Alzheimer's disease, where imaging modalities and genetic modality are likely to link through a highly non-linearly manner. To capture the complex interactions among modalities, we thus propose a novel framework to fuse multiple data modalities through deep matrix factorization. Assume we have t data modalities $X_1 \in \mathbb{R}^{n\times d_1}, \ldots, X_t \in \mathbb{R}^{n\times d_t}$ describing different views of the same set of n samples. We use a deep neural network $g_\theta(.)$ parameterized θ to factorize each modality, i.e., $X_i \approx U g_{\theta_i}(V_i)$, where in

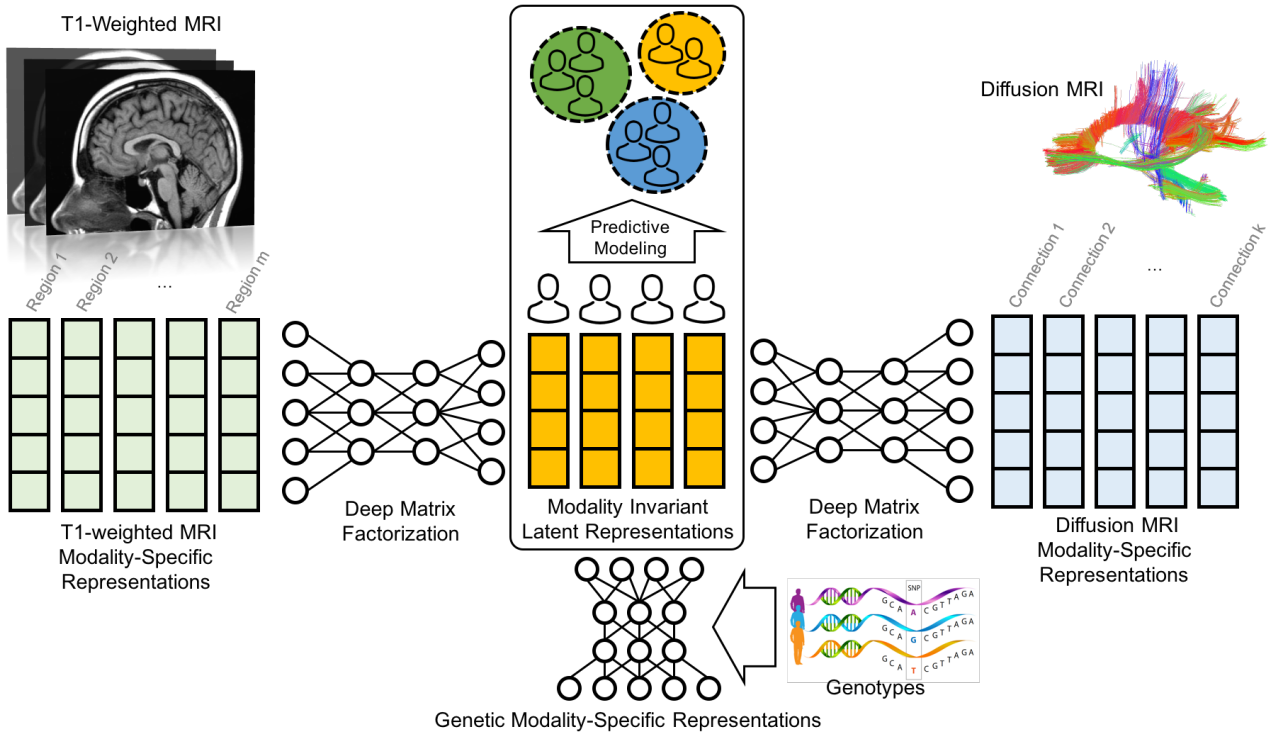

Figure 1: Illustration of proposed collective deep matrix factorization (CDMF) framework. In this example, CDMF fuses information from three modalities: T1 weighted MRI, diffusion MRI, and genotypes (SNPs) to learn a modality invariant latent representation, to perform predictive modeling.

this paper we use a structured deep neural network with k layers:

$$g_{\theta_i}(V_i) = f(W_{(k,i)}f(W_{(k-1,i)}f(\ldots, f(W_{(1,i)}V_i)),$$

where $W_{(j,i)}$ is the network weight at the j-th layer, $\theta_i = \{W_{(k,i)}, W_{(k-1,i)}, \ldots, W_{(1,i)}\}$ collectively denotes network weights, and f is a non-linear activation function. The deep network serves as a highly non-linear mapping between input matrix X_i and U, and projects the latent representations non-linearly to the same latent space. We call this $g_{\theta_i}(V_i)$ modality specific component for i-th modality. We can thus perform collective deep matrix factorization (CDMF) to associate multiple data modalities:

$$\min_{U,\{V_i,\theta_i\}_{i=1}^t} \sum_{i=1}^t d(X_i, Ug_{\theta_i}(V_i)) \text{ s.t. } U \in \mathcal{S}_0, V_i \in \mathcal{S}_i, \forall i.$$

We would like to highlight one property of collective deep matrix factorization that modality invariant component/representation can have different dimensions from modality components, i.e., U and V can be different, and V in different modalities can also be different. This flexibility is desired especially when different modalities contain different amount of information, and thus the optimal latent representations may have different dimensions. We also note that one way to control the complexity of networks under multiple modalities is to enforce shared network structures, i.e., $\{g_{\theta_i}\}$ have the same architecture and share the same parameter values, except for the last layer.

In many applications, our ultimate goal is to build predictive models from multi-modal analysis. To achieve this, we can integrate predictive modeling and collective deep matrix factorization during learning, such that predictive modeling uses latent representations learned from collective deep matrix factorization as input features. Assume that we are given supervision information $\{y_1, \ldots, y_n\}$ for the n subjects, and a linear model for the prediction task $h(U; \mathbf{w}) = U\mathbf{w}$ (with a dummy variable to include bias). Given a latent representation U_j (i.e. the j-th row of U matrix) for the j-th subject and its corresponding label y_j, we use a proper loss function $\ell(h(U_j; \mathbf{w}), y_j)$ (e.g., logistic loss for classification and least squares for regression). The proposed supervised CDMF formulation is thus given by:

$$\min_{\mathbf{w},U,\{V_i,\theta_i\}_{i=1}^t} \sum_{j=1}^n \ell(h(U_j; \mathbf{w}), y_j) + \sum_{i=1}^t \alpha_i d(X_i, Ug_{\theta_i}(V_i))$$
$$\text{s.t. } U \in \mathcal{S}_0, V_i \in \mathcal{S}_i, \forall i, \quad (2)$$

where α_i is a tunable parameter to control knowledge fusion proportion of the i-th modality, specifying how much that the modality influences the learning of the modality invariant component. When α_i is large, a less reconstruction error for this modality will be achieved when minimizing overall loss, and therefore the learned representation U contains more information of this modality, and vice versa. Figure 1 illustrates the proposed framework fusing three modalities: dMRI, T1 MRI and genotypes (SNPs).

Optimization and initialization. The formulation can be solved efficiently by TensorFlow [1]. However, since the objective in (2) is highly non-convex and gradient algorithms may easily trapped in local optima, a good initialization is important for training the

network. In this work, we propose to iteratively apply linear matrix factorizations in the original data matrix, and use linear and hierarchical matrix factors to initialize the deep neural networks. As such, the initialization is similar to a valid linear matrix factorization, and the algorithm iteratively explore non-linear effects within linear latent spaces and capture non-linearity in the network during learning process. Technically we can choose arbitrary linear factorization methods in (1) for initialization, however, we find in our experiments that singular vectors given by iterative singular value decomposition (SVD) usually provide decent models that outperform other factorization methods. This may due to fact that orthogonal basis obtained by SVD characterize the optimal linear subspace of the data matrix.

Handling modalities with missing subjects. In many applications especially medical cases, some data modalities may not be available to all samples. For example, some subjects did not participate the genetic study and thus lack genotype information. Besides, in the first stage of ADNI study there are no diffusion MRI imaging available, leading to structured missing patterns in the dataset [46]. Since $\{X_i\}$ involve different sets of subjects, such missing modalities will cause dimension problems in U, and thus the modalities cannot be projected to the same U. One way to overcome this issue is to discard all the subjects with missing modalities and make the dimensions consistent across modalities. However, this approach will significantly reduce the number of samples and thus compromise the predictive performance. We therefore extend the proposed formulation to deal with it. We define an indicator matrix for each modality, where for the i-th modality it is denoted by $I_i \in \mathbb{R}^{n\times n}$, whose j-th row is given by:

$$(I_i)_j = \begin{cases} \mathbf{0} & \text{if the this modality is missing for } j\text{-th subject} \\ \mathbf{e}_j & \text{otherwise} \end{cases},$$

where $\mathbf{e}_j \in \mathbb{R}^n$ is n-dimensional standard basis with only j-th entry as 1. The revised formulation is given by:

$$\min_{\mathbf{w}, U, \{V_i, \theta_i\}_{i=1}^t} \sum_{j=1}^n \ell(h(U_j; \mathbf{w}), y_j) + \sum_{i=1}^t \alpha_i d(\hat{X}_i, I_i U g_{\theta_i}(V_i)) \quad \text{s.t. } U \in \mathcal{S}_0, V_i \in \mathcal{S}_i, \forall i, \tag{3}$$

where $\hat{X}_i$ is an augmented data matrix, whose j-th row is given by:

$$\hat{X}_i^j = \begin{cases} \mathbf{0} & \text{if this subject lacks of } i\text{-th modality} \\ X_i^j & \text{(original features) otherwise} \end{cases}.$$

By multiplying indicators and replacing X_i by $\hat{X}_i$, the corresponding rows of subjects with missing modality will be 0 for this modality, which has no effect on loss. This approach would ensure that we use all the information available during the learning.

Application in Disease Modeling. Even though the proposed CDMF framework can be used in various data mining applications, here we emphasize on its advantages in our specific disease modeling problem. The goal of MCI diagnosis is to differentiate between MCI subjects and normal cognitive (NC) subjects, which is a classification problem. We thus use CDMF in Eq. (3) with a logistic loss, in which knowledge from different modalities is fused in a supervised manner such that only the part that is more relevant to group difference of MCI and NC will be fused to the latent representation U, which in turn can improve prediction. This property is important for our multi-modality disease modeling since the modalities may contain knowledge that is not relevant to the desired learning task. Without proper guidance, the irrelevant knowledge may negatively impact the representation leading to suboptimal predictive performance. For example, brain imaging may contain information of other inherited brain diseases or aging properties, likewise for genetic data. If the fusion process is carried out in an unsupervised manner, we may not obtain a U that is most informative regarding the progression of MCI.

Association study of multiple modalities. The interactions between latent representations are of great interests in the community (e.g., generate predictions in the recommender system), and can reveal important insights into how different modalities are connected to each other. Although it is straight forward in linear case that we can use inner products $\mathbf{u}^i(\mathbf{v}^j)^T$, we cannot directly compute this way in CDMF since the modalities are connected through non-linear networks. Instead, we can use the following transformed latent factors:

$$\tilde{V}_i = f(W_{(k,i)} f(W_{(k-1,i)} f(\ldots, f(W_{(1,i)} V_i)), \tag{4}$$

which is a mapping matrix that contains the modality specific information of the corresponding modality. All the columns of this matrix form the specific feature space of this modality. Hence, we can calculate the association of any features between any two modalities using the transformed latent factors $\tilde{V}_i$. Let $C_{i,j}(m,n)$ denote the cosine similarity between the m-th column from $\tilde{V}_i$ and the n-th column from $\tilde{V}_j$. When $C_{i,j}(m,n)$ is large, the m-th feature of i-th modality is highly related with the n-th feature of j-th modality and a small $C_{i,j}(m,n)$ indicates the association between those features is weak. This provides a novel tool to study the imaging genetics, identifying how genotypes influence brain structures under specific tasks (e.g., MCI prediction in our case).

4 EXPERIMENT

4.1 Dataset and features

Data from two stages of ADNI are used in this study: ADNI1 and ADNI2. Detail demographic characteristics and missing data information are listed in Table 1. Whole genome sequencing (WGS) SNPs are provided by ADNI and used as genetic modality in our study. For MRI, ADNI1 participants are scanned by 1.5T or 3T MRI scanner while all ADNI2 participants are scanned by 3T MRI scanner [1]. FreeSurfer V5.3 is adopted to extract 333 measures include the area, thickness, cortical volume, subcortical volume and white matter volume from T1 MRI to form T1 MRI modality. For dMRI, we first parcellate the brain into 113 cortical and subcortical region-of-interests (ROIs) according to the Harvard Oxford Cortical and subcortical Probabilistic Atlas [13]. Then we reconstruct the whole-brain tractography using an ODF-based probabilistic approach: PICo[12]. Finally, a brain network is generated in which the nodes indicate ROIs and the edges are determined by the proportion of fibers intersecting with each pair of ROIs. As such, each brain network is a 113×113 symmetric matrix with 6328 distinct edges. These 6328 edges are used as the feature variables for dMRI modality.

[1] http://adni.loni.usc.edu/data-samples/mri/

Table 1: Demographic information of subjects

ADNI1 Cohort	NC	MCI	Total	**ADNI2 Cohort**	NC	MCI	Total
Age	75.84±4.95	74.48±7.48	75.17±6.68	Age	69.36 ± 15.40	71.68 ± 9.93	70.96 ± 11.89
Sex	115M/108F	247M/138F	362M/246F	Sex	22M/28F	71M/41F	93M/69F
total subjects	223	385	608	total subjects	50	112	162
Subjects with dMRI	0	0	0	Subjects with dMRI	50	112	162
Subjects with T1 MRI	223	385	608	Subjects with T1 MRI	50	112	162
Subjects with genotype	202	348	550	Subjects with genotype	27	82	109

4.2 Data preprocessing

Imaging modalities preprocessing. ADNI1 and ADNI2 use different scanner protocol which may introduce biases for the datasets. Hence, we decide to harmonize the cohorts by removing this cohort effect. We create an indicator variable to differentiate ADNI1 and ADNI2 with 1 for all subjects from ADNI1 and -1 for all subjects from ADNI2. In addition, age and sex are common confounders biasing the analysis. In this study, generalized linear regression approach [31] is used to remove all confounders including age, sex and cohort index. It assumes each observed variable is linearly dependent on the confounder variables and fitting a generalized linear model can remove confounders' effect. Denote the observed variable of variable X as X^{obs} and the original variable as X^{ori}. The linear dependence of X^{obs} and X^{ori} is:

$$X^{obs} = w_1 \cdot age + w_2 \cdot sex + w_3 \cdot cohort + X^{ori},$$

where w_1, w_2, w_3 are coefficients of confounders. Let (w_1, w_2, w_3) be $\boldsymbol{w}$ and $(age_i, sex_i, cohort_i)$ be t_i, where i denotes the i-th subject. Coefficients can be obtained by solving a linear regression:

$$\boldsymbol{w}^* = \min_{\boldsymbol{w}} \sum_{i=1}^{n} (\boldsymbol{w}^T t_i - X_i^{obs})^2. \quad (5)$$

After solving Eq. (5), the original feature variable is given by:

$$X^{ori} = X^{obs} - (w_1 \cdot age + w_2 \cdot sex + w_3 \cdot cohort).$$

We apply this on both T1 MRI data and dMRI data and will only use X^{ori} in the downstream experiments.

Genetic modality preprocessing. Genetic data is preprocessed by standard quality control using PLINK[2] and then impute using MaCH[3]. SNPs with minor allele frequency (MAF) less than 5% or missing values greater than 5% are discarded. Subjects with missing values greater than 10% at all SNPs are removed. Finally, 659 subjects with reading values on $6,566,154$ SNPs are attained.

In order to extract more relevant features, we apply genome-wide association study (GWAS) on our data. In detail, we regress patient state NL/MCI on each SNP using logistic regression, with p-value generated and adjusted to $-\log_{10}$ scale. Larger adjusted p-value indicates strong association between response and the marker. Figure 2 shows SNPs with adjust p value greater than 2 on each chromosome. SNPs on chromosome 19 have stronger association with MCI than others, suggesting crucial effects of this chromosome on the Alzheimer's deterioration. Finally, the top 200 significant SNPs for each iteration are retained as features for our downstream analysis. Since SNPs are categorical, i.e. $\{0, 1, 2\}$, we use the one-hot coding to be the feature representation. Hence, the final feature dimension for genetic modality is 600.

[2]http://pngu.mgh.harvard.edu/ purcell/plink/
[3]http://csg.sph.umich.edu/abecasis/MaCH/

4.3 Predict performance

Comp. #	Shallow collective matrix factorization		
	linear	sigmoid	square
30	0.529 ± 0.080	0.616 ± 0.102	0.564 ± 0.011
50	0.587 ± 0.069	0.593 ± 0.120	0.718 ± 0.076
70	0.610 ± 0.079	0.644 ± 0.075	0.659 ± 0.161
90	0.526 ± 0.065	0.597 ± 0.086	0.634 ± 0.097
110	0.656 ± 0.089	0.681 ± 0.116	0.658 ± 0.106
130	0.561 ± 0.024	0.613 ± 0.105	0.668 ± 0.127
Comp. #	Deep collective matrix factorization		
	linear	sigmoid	square
30	0.519 ± 0.099	0.653 ± 0.139	0.719 ± 0.142
50	0.594 ± 0.151	0.646 ± 0.078	0.693 ± 0.100
70	0.573 ± 0.135	0.593 ± 0.165	0.758 ± 0.115
90	0.519 ± 0.093	0.610 ± 0.146	**0.805 ± 0.073**
110	0.558 ± 0.083	0.542 ± 0.048	0.726 ± 0.027
130	0.553 ± 0.124	0.544 ± 0.110	0.679 ± 0.152
Comp. #	Other deep multi-modality methods		
	DCCA	DCCAE	DNN
30	0.770 ± 0.065	0.723 ± 0.031	0.617 ± 0.143
50	0.722 ± 0.088	0.743 ± 0.094	0.604 ± 0.026
70	0.689 ± 0.134	0.780 ± 0.054	0.560 ± 0.111
90	0.684 ± 0.089	0.703 ± 0.042	0.579 ± 0.068
110	*	0.735 ± 0.135	0.627 ± 0.165
130	*	0.699 ± 0.089	0.689 ± 0.131

* means not applicable due to the algorithm design.

Table 2: Prediction performance of different models using ADNI2's T1 MRI and dMRI in terms of AUC. With an appropriate activation function and components' number, our method outperforms than all other methods.

In this section, we evaluate the performance of our method and compare with other methods using ADNI dataset. The distance metric $d(X, Y)$ we used in the following experiments is $\|X - Y\|_F^2$. We perform experiments on three different settings.

In the first setting, only ADNI2 dataset and its two modalities: T1 MRI and dMRI are covered. In this setting, no modality has missing subjects. We randomly select 90% subjects as the training set and 10% subjects as the testing set. Our main assumption is deep matrix factorization can extract high-level nonlinear features to improve diagnosis performance. In order to prove it, we compare deep models with shallow models, i.e. one layer matrix factorization, and compare nonlinear models with linear models. Two main nonlinear functions are used in our experiments: $sigmoid(x)$ and x^2. In deep models, we focus on those with two hidden layers.

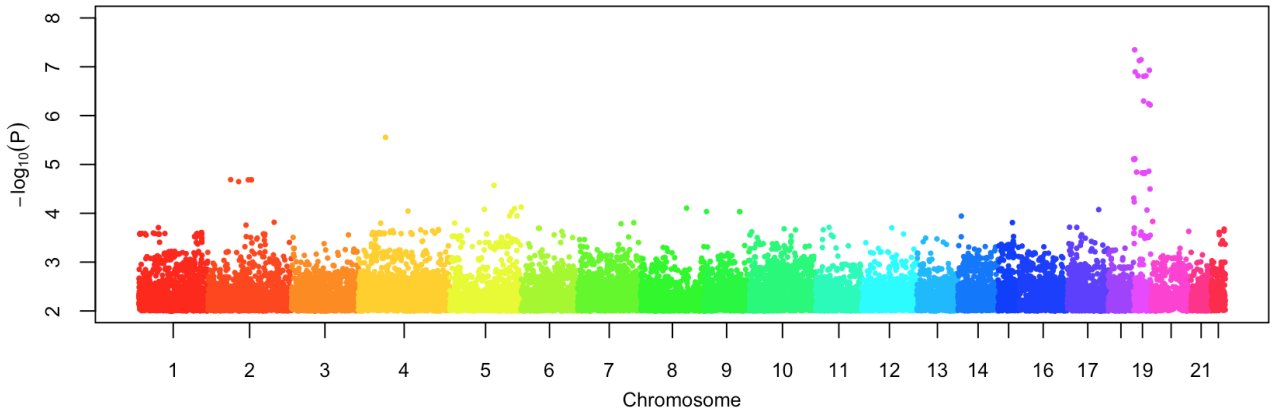

Figure 2: Manhattan plot for SNPs with adjusted p value greater than 2. Colors indicate different chromosomes.

Comp. #	Shallow collective matrix factorization		
	linear	sigmoid	square
30	0.702 ± 0.019	0.672 ± 0.137	0.708 ± 0.024
50	0.749 ± 0.052	0.793 ± 0.034	0.742 ± 0.063
70	0.743 ± 0.063	0.696 ± 0.037	0.747 ± 0.061
90	0.754 ± 0.046	0.756 ± 0.059	0.749 ± 0.049
110	0.791 ± 0.027	0.798 ± 0.058	0.786 ± 0.032
130	0.671 ± 0.049	0.652 ± 0.058	0.679 ± 0.048
Comp. #	Deep collective matrix factorization		
	linear	sigmoid	square
30	0.634 ± 0.065	0.665 ± 0.044	0.627 ± 0.768
50	0.701 ± 0.064	0.735 ± 0.061	0.681 ± 0.039
70	0.778 ± 0.059	0.749 ± 0.011	0.784 ± 0.055
90	0.775 ± 0.063	0.801 ± 0.023	**0.821 ± 0.015**
110	0.806 ± 0.049	0.792 ± 0.031	0.800 ± 0.032
130	0.717 ± 0.037	0.705 ± 0.049	0.759 ± 0.044
Comp. #	Other deep multi-modality methods		
	DCCA	DCCAE	DNN
30	0.801 ± 0.101	0.737 ± 0.063	0.758 ± 0.098
50	0.732 ± 0.041	0.753 ± 0.014	0.767 ± 0.069
70	0.788 ± 0.084	0.813 ± 0.047	0.756 ± 0.087
90	0.746 ± 0.159	0.750 ± 0.124	0.757 ± 0.078
110	0.759 ± 0.151	0.780 ± 0.058	0.754 ± 0.070
130	0.739 ± 0.183	0.774 ± 0.074	0.754 ± 0.056

Table 3: Prediction performance of different models using ADNI2 and ADNI1's T1 MRI and dMRI in terms of AUC. Although dMRI modality lacks of a large number of subjects, performance is still improved a lot compared with that only uses ADNI2 data.

After some preliminary experiments, we fix the first layer's components to be 162, i.e. $V_j \in R^{162\times d_j}$ for $j = 1, 2, ..., t$ and vary second layer's components from 30 to 130, i.e. $W_{1,j} \in R^{r\times 162}$ where $r \in \{30, 50, 70, 90, 110, 130\}$. Hence, $U \in R^{n_j\times r}$. How the new features' dimension affects performance can be traced by varying r. We report average area under ROC curve (AUC) over three iterations in Table 2. We implemented the proposed model using TensorFlow [1]. All the experiments were run on GT1080 or Titan X. It takes approximately 3 minutes to train one model.

When using x^2 as activation function and setting components number to be 90, our model outperforms all other models. We observe when the activation function is inappropriate, i.e, $sigmoid(x)$ for our case, the AUC is very low. Hence, choosing a suitable activation function is very important. Only certain nonlinear functions can correctly fit this dataset and extract the desired features. Also, we find the number of components is crucial for all different models. An inappropriate number of components will reduce the performance drastically. When the number of components is too small, new feature representation is not rich enough to capture the complex hidden information. But when this number becomes too large, they contain too many redundant features. Since sample size is not large enough, it causes overfitting and reduces testing performance. We also compare our method with three state-of-the-art multi-modality learning algorithms: DCCA, DCCAE and deep neural network. Since training sample size is 90, when the components number of new feature representation is larger than 90, DCCA's code[4] reports error. Hence, we set it to be $\{30, 50, 70, 90\}$ for DCCA. The deep neural network has two parts. The first part is used to remove modality specific information. It has two two-layer sub-networks corresponding to two modalities. The first layer is the input layer. To make the network consistent. The second layer contains 162 neurons for each sub-network. The outputs of two sub-networks are concatenate to a vector and used as the input of the second part of the whole network to fuse knowledge and implement classification tasks. The second part has three layers. The first layer is the input layer where the output of the first part is fed. The second layer contains $\{30, 50, 70, 90, 110, 130\}$ units. The third layer is a logistic regression layer. To compare with our model, the two parts are jointly trained. The results are reported in the last three columns in Table 2. Our method outperforms all baselines.

In the second setting, we include all ADNI1 subjects' imaging data into the training set. Compared with the first setting, dMRI

[4]http://ttic.uchicago.edu/ wwang5/dccae.html

Components #	Shallow collective matrix factorization			Deep collective matrix factorization			DNN
	linear	sigmoid	square	linear	sigmoid	square	
30	0.684 ± 0.051	0.658 ± 0.039	0.766 ± 0.115	0.632 ± 0.019	0.665 ± 0.042	0.670 ± 0.052	0.674 ± 0.114
50	0.767 ± 0.019	0.772 ± 0.032	0.818 ± 0.076	0.707 ± 0.054	0.737 ± 0.064	0.719 ± 0.073	0.666 ± 0.108
70	0.763 ± 0.059	0.759 ± 0.020	0.797 ± 0.049	0.781 ± 0.065	0.750 ± 0.010	0.799 ± 0.040	0.669 ± 0.119
90	0.772 ± 0.070	0.775 ± 0.030	0.767 ± 0.081	0.784 ± 0.071	0.797 ± 0.019	**0.852 ± 0.018**	0.667 ± 0.090
110	0.822 ± 0.018	0.795 ± 0.005	0.803 ± 0.014	0.811 ± 0.047	0.782 ± 0.030	0.779 ± 0.008	0.656 ± 0.080
130	0.702 ± 0.067	0.669 ± 0.055	0.689 ± 0.071	0.728 ± 0.048	0.705 ± 0.055	0.725 ± 0.105	0.671 ± 0.098

Table 4: Prediction performance of fusing genetic knowledge and imaging knowledge using ADNI1 and ADNI2 in terms of AUC. Genetic modality can be successfully integrated with imaging modalities.

modality has a lot of missing subjects in this setting. Also, this setting's training sample size is much larger than the previous one. In order to compare the performance of these two settings, the testing data set and all the other model settings are the same as in the first setting. Since DNN, DCCA and DCCAE cannot deal with modality with missing subjects, we fill all the missing values with the mean over all available samples for each modality. Average AUC is reported in Table 3 for all models and similar trends are observed in these results with those in the first setting. Moreover, we find under the same experiment settings, almost all models' performance is higher than that of the previous one. It shows our extended formulation can successfully deal with modality with missing subjects and leverage partial knowledge in this modality to greatly improve overall performance.

In the last setting, we include genetic modality as the third modality and fuse genetic knowledge and imaging knowledge to improve diagnosis performance. We preform GWAS on each iteration's training set to select SNPs involved in our experiment. To compare with the second setting, all the model settings are the same as in previous settings. Average AUC is reported in Table 4. Since DCCA and DCCAE cannot deal with three modalities, we only use DNN as baseline. With the same training sample size, DNN's performance is much worse than that of previous setting, which implies concatenating all the output of each sub-network as fusion method does not work for this case. That is because features from the genetic modality are discrete and the matrix is very sparse, while features for two imaging modalities are continues and the matrices are extremely dense. They have different statistical properties. However, for our method, the performance for this setting is much better than that of the second setting, which implies genetic modality can be successfully integrated with imaging modalities by our method even though the modalities are radically different.

At last, we report sparse logistic regression results on each single modality as single modality baselines. The results are shown in Table 5. Experiments on ADNI2 dataset have the same training testing splitting method as the first setting and experiments on ADNI1 + ADNI2 dataset have the same splitting way as the second setting. We see single modality's average AUC is lower than the highest AUC in all three settings. Hence, only by fusing knowledge from different modalities can we achieve descent performance.

	ADNI2		
	T1 MRI	dMRI	SNPs
AUC	0.71 ± 0.04	0.63 ± 0.07	0.63 ± 0.14
	ADNI1+ADNI2		
	T1 MRI	dMRI	SNPs
AUC	0.72 ± 0.06	-	0.67 ± 0.27

Table 5: Results of applying sparse logistic regression on each single modality in terms of AUC. ADNI1 study did not collect dMRI.

4.4 Effects of knowledge fusion parameters

Knowledge fusion parameters control how much knowledge a modality is fused into modality invariant term. In this section, we show how these parameters affect performance. Let $\alpha_1, \alpha_2, \alpha_3$ be the parameters to control knowledge fusion of dMRI, T1 MRI and SNPs respectively. The training set and testing set are split in the same way as the third setting in the last section. We focus on deep model with 2 hidden layers, with x^2 as activation function. The components of the first layer and the second layer is 162 and 90 respectively.

We first fix α_1 and α_2 to be 1 and vary α_3 to see how α_3 affects performance. The results is shown in Figure 4 in blue line. We see when we increase α_3, the performance first increases slightly. But when α_3 is larger than 0.1, the performance decreases very fast if we continue increasing it. That is because genetic modality is noisier than imaging modalities. With a small α_3, i.e. 0.1, this model can tolerant a larger reconstruction error for genetic modality. Hence, the model is robust to the noise in genetic modality. When α_3 becomes larger, the reconstruction error of genetic modality must be small in order to achieve a low total loss. More noise distorts U, which reduces the performance. But when α_3 is too small, some useful knowledge of this modality cannot all be fused to U, which also reduces the performance. Hence, only with a suitable fusion parameter can the model correctly fuses all the useful knowledge of genetic modality. Next, we fix α_3, α_1 to be 0.1 and 1 receptively and vary α_2 to see how α_2 affects performance. We also fix α_3, α_2 to be 0.1 and 1 respectively and vary α_1 to see the effects of changing α_1. The results are shown in Figure 4 in green line and red line. These two are very similar to each other since they both control knowledge fusion of imaging modalities. We see when α_1 and α_2 reach 1, the performance reaches the highest. Hence, imaging modalities need to contribute more knowledge to U than genetic modality to make a better performance.

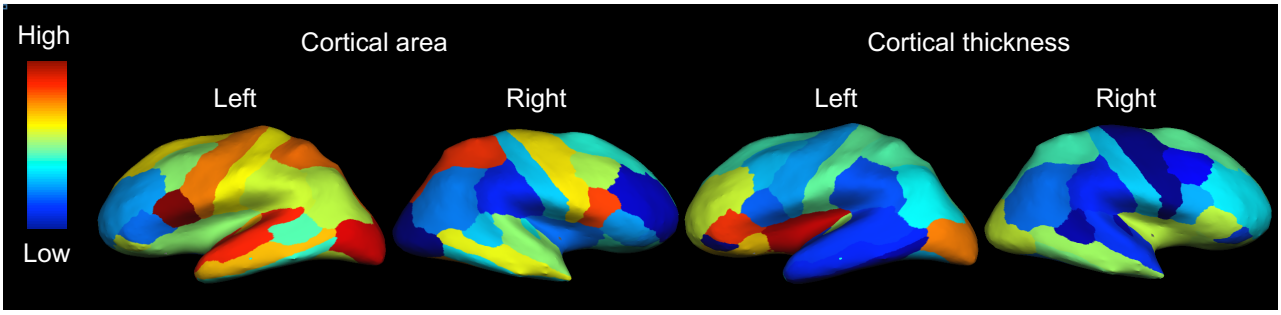

Figure 3: Brain maps of the significance level at each ROI for the most associated SNP within that ROI.

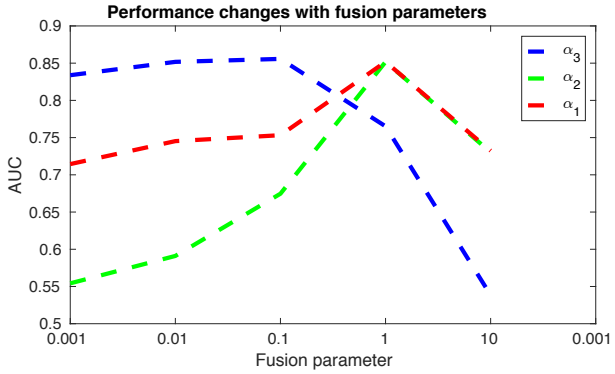

Figure 4: Testing performance with varying α parameters.

4.5 Imaging-genetics association

In this section, we present imaging-genetics association uncovered by modality specific components. We compute the association between SNPs with cortical thickness and area on 68 ROIs. This association indicates how significant a brain imaging feature is associated with a SNP under the task of predicting MCI. In Figure 3, we show the map of the significance level at each ROI for the most associated SNP within that ROI. The first two figures are based on cortical area features for left and right brain respectively and the last two figures are for cortical thickness features. Warmer colors represent stronger association and cooler colors indicate the opposite. Our results show that there are some cluster patterns which indicate those ROIs are highly related to each other in respect of MCI. Top 6 significant T1 MRI features are: right cuneus thickness, right parahippocampal area, right posterior cingulate thickness, left pars opercularis area, left cuneus thickness and right frontal pole thickness. Among those features, cuneus thickness, posterior cingulate thickness, frontal pole thickness and parahippocampal region are identified significantly associated with MCI [10, 14, 18, 33]. The SNPs most related to these 6 features are: rs10414043, rs429358, rs429358, rs8141950, rs11178933, rs10414043 respectively. All the SNPs except rs8141950 are located at Chromosome19 which has been identified to be highly associated with MCI and AD [11, 27]. Especially, rs429358 locates in the fourth exon of the APOE gene [21] in Chromosome19, which has been extensively reported as the genetic risk factor for the late-onset of AD. rs8141950, located on Chromosome22, has also been found to be closely related to AD [2]. This shows that our method can correctly uncover imaging-genetic association in respect of MCI. This association can be used to analyze how the genotype influences brain structures and provide a potential way to explore the mechanism behind MCI and AD.

5 CONCLUSIONS

In this paper, we proposed collective deep matrix factorization to fuse knowledge from different modalities. Specifically, we build uniform nonlinear hierarchical deep matrix factorization framework across different modalities which decomposes each modality into a modality specific component and a modality invariant component that serves as a learned feature representation. We also add supervision on the modality invariant component to guide the learning process. The proposed method can exploit complicated non-linear interactions among different modalities and learn a feature representation which is compact and more relevant to our predictive problem. Also, the modality specific term can be used to uncover complicated imaging-genetic associations. We perform extensive experiments on ADNI dataset and show the proposed method significantly improves predictive performance. We also show some imaging-genetic association which can benefit future research. We plan to extend this framework into the multi-task learning setting [47] to simultaneously predict multiple targets of interests.

ACKNOWLEDGMENTS

This research is supported in part by National Science Foundation under Grant IIS-1565596 (JZ), IIS-1615597 (JZ), IIS-1615035 (SJ), and DBI-1641223 (SJ), the Office of Naval Research under grant number N00014-17-1-2265 (JZ), N00014-14-1-0631 (JZ), and National Institutes of Health under Grant U54 EB020403 (PT). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the GPU cards used for this research.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, and others. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
[2] H Akalın, Yahya Karaman, H Demirtaş, N İmamoğlu, Yusuf Özkul, and others. 2005. Evaluation of the nucleolar organizer regions in Alzheimer's disease. *Gerontology* 51, 5 (2005), 297–301.
[3] Zeynep Akata, Christian Thurau, and Christian Bauckhage. 2011. Non-negative matrix factorization in multimodality data for segmentation and label prediction. In *16th Computer vision winter workshop*.
[4] Alzheimer's Association. 2013. 2013 Alzheimer's disease facts and figures. (2013).

[5] alz.org. 2017. Mild Cognitive Impairment. Retrieved at http://www.alz.org/dementia/mild-cognitive-impairment-mci.asp. (2017).

[6] Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu. 2013. Deep Canonical Correlation Analysis.. In *ICML*. 1247–1255.

[7] Lars Bertram and Rudolph E Tanzi. 2008. Thirty years of Alzheimer's disease genetics: the implications of systematic meta-analyses. *Nature Reviews Neuroscience* 9, 10 (2008), 768–778.

[8] Emmanuel Candes and Benjamin Recht. 2012. Exact matrix completion via convex optimization. *Commun. ACM* 55, 6 (2012), 111–119.

[9] Shiyu Chang, Guo-Jun Qi, Charu C Aggarwal, Jiayu Zhou, Meng Wang, and Thomas S Huang. 2014. Factorized similarity learning in networks. In *ICDM*. IEEE, 60–69.

[10] Yu-Ling Chang, Mark W Jacobson, Christine Fennema-Notestine, Donald J Hagler, Robin G Jennings, Anders M Dale, Linda K McEvoy, Alzheimer's Disease Neuroimaging Initiative, and others. 2010. Level of executive function influences verbal memory in amnestic mild cognitive impairment and predicts prefrontal and posterior cingulate thickness. *Cerebral Cortex* 20, 6 (2010), 1305–1313.

[11] EH Corder, AM Saunders, WJ Strittmatter, DE Schmechel, PC Gaskell, GWet al Small, AD Roses, JL Haines, and Margaret A Pericak-Vance. 1993. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261, 5123 (1993), 921–923.

[12] Maxime Descoteaux, Rachid Deriche, Thomas R Knosche, and Alfred Anwander. 2009. Deterministic and probabilistic tractography based on complex fibre orientation distributions. *IEEE transactions on medical imaging* 28, 2 (2009), 269–286.

[13] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, and others. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 31, 3 (2006), 968–980.

[14] DP Devanand, Ravi Bansal, Jun Liu, Xuejun Hao, Gnanavalli Pradhaban, and Bradley S Peterson. 2012. MRI hippocampal and entorhinal cortex mapping in predicting conversion to Alzheimer's disease. *Neuroimage* 60, 3 (2012), 1622–1629.

[15] Robin D Dowell, Owen Ryan, An Jansen, Doris Cheung, Sudeep Agarwala, Timothy Danford, Douglas A Bernstein, P Alexander Rolfe, Lawrence E Heisler, Brian Chin, and others. 2010. Genotype to phenotype: a complex problem. *Science* 328, 5977 (2010), 469–469.

[16] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. 2015. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *WWW*. ACM, 278–288.

[17] David C Glahn, Paul M Thompson, and John Blangero. 2007. Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function. *Human brain mapping* 28, 6 (2007), 488–501.

[18] Päivi Hartikainen, Janne Räsänen, Valtteri Julkunen, Eini Niskanen, Merja Hallikainen, Miia Kivipelto, Ritva Vanninen, Anne M Remes, and Hilkka Soininen. 2012. Cortical thickness in frontotemporal dementia, mild cognitive impairment, and Alzheimer's disease. *Journal of Alzheimer's Disease* 30, 4 (2012), 857–874.

[19] Clifford R Jack, Matt A Bernstein, Nick C Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J Britson, Jennifer L Whitwell, Chadwick Ward, and others. 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of magnetic resonance imaging* 27, 4 (2008), 685–691.

[20] Meina Kan, Shiguang Shan, and Xilin Chen. 2016. Multi-view deep network for cross-view classification. In *CVPR*. 4847–4855.

[21] Jungsu Kim, Jacob M Basak, and David M Holtzman. 2009. The role of apolipoprotein E in Alzheimer's disease. *Neuron* 63, 3 (2009), 287–303.

[22] Marius Kloft, Ulf Brefeld, Pavel Laskov, Klaus-Robert Müller, Alexander Zien, and Sören Sonnenburg. 2009. Efficient and accurate lp-norm multiple kernel learning. In *NIPS*. 997–1005.

[23] Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer* 42, 8 (2009).

[24] Michael Krawczak, Susanna Nikolaus, Huberta von Eberstein, Peter JP Croucher, Nour Eddine El Mokhtari, and Stefan Schreiber. 2006. PopGen: population-based recruitment of patients and controls for the analysis of complex genotype-phenotype relationships. *Public Health Genomics* 9, 1 (2006), 55–61.

[25] Daniel D Lee and H Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (1999), 788–791.

[26] Daniel D Lee and H Sebastian Seung. 2001. Algorithms for non-negative matrix factorization. In *NIPS*. 556–562.

[27] Chia-Chan Liu, Takahisa Kanekiyo, Huaxi Xu, and Guojun Bu. 2013. Apolipoprotein E and Alzheimer disease: risk, mechanisms and therapy. *Nature Reviews Neurology* 9, 2 (2013), 106–118.

[28] Xiaoxiao Liu, Marc Niethammer, Roland Kwitt, Nikhil Singh, Matt McCormick, and Stephen Aylward. 2015. Low-rank atlas image analyses in the presence of pathologies. *IEEE transactions on medical imaging* 34, 12 (2015), 2583–2591.

[29] David Meunier, Renaud Lambiotte, Alex Fornito, Karen D Ersche, and Edward T Bullmore. 2010. Hierarchical modularity in human brain functional networks. *Hierarchy and dynamics in neural networks* 1, 2 (2010).

[30] José Luis Molinuevo, Pablo Ripolles, Marta Simó, Albert Lladó, Jaume Olives, Mircea Balasa, Anna Antonell, Antoni Rodriguez-Fornells, and Lorena Rami. 2014. White matter changes in preclinical Alzheimer's disease: a magnetic resonance imaging-diffusion tensor imaging study on cognitively normal older people with positive amyloid β protein 42 levels. *Neurobiology of aging* 35, 12 (2014), 2671–2680.

[31] John A Nelder and R Jacob Baker. 1972. Generalized linear models. *Encyclopedia of statistical sciences* (1972).

[32] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *ICML*. 689–696.

[33] Eini Niskanen, Mervi Könönen, Sara Määttä, Merja Hallikainen, Miia Kivipelto, Silvia Casarotto, Marcello Massimini, Ritva Vanninen, Esa Mervaala, Jari Karhu, and others. 2011. New insights into Alzheimer's disease progression: a combined TMS and structural MRI study. *PLoS One* 6, 10 (2011), e26113.

[34] Ronald C Petersen, Rachelle Doody, Alexander Kurz, Richard C Mohs, John C Morris, Peter V Rabins, Karen Ritchie, Martin Rossor, Leon Thal, and Bengt Winblad. 2001. Current concepts in mild cognitive impairment. *Archives of neurology* 58, 12 (2001), 1985–1992.

[35] Tara N Sainath, Brian Kingsbury, Vikas Sindhwani, Ebru Arisoy, and Bhuvana Ramabhadran. 2013. Low-rank matrix factorization for deep neural network training with high-dimensional output targets. In *ICASP*. IEEE, 6655–6659.

[36] Dennis J Selkoe. 1996. Amyloid β-protein and the genetics of Alzheimer's disease. *Journal of Biological Chemistry* 271, 31 (1996), 18295–18298.

[37] Hyun Ah Song, Bo-Kyeong Kim, Thanh Luong Xuan, and Soo-Young Lee. 2015. Hierarchical feature extraction by multi-layer non-negative matrix factorization network for classification task. *Neurocomputing* 165 (2015), 63–74.

[38] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjorn W Schuller. 2015. A deep matrix factorization method for learning attribute representations. (2015).

[39] P Vemuri, HJ Wiste, SD Weigand, David S Knopman, JQ Trojanowski, LM Shaw, Matthew A Bernstein, PS Aisen, M Weiner, Ronald Carl Petersen, and others. 2010. Serial MRI and CSF biomarkers in normal aging, MCI, and AD. *Neurology* 75, 2 (2010), 143–151.

[40] Qi Wang, Liang Zhan, Paul M Thompson, Hiroko H Dodge, and Jiayu Zhou. 2016. Discriminative fusion of multiple brain networks for early mild cognitive impairment detection. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 568–572.

[41] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A Bilmes. 2015. On Deep Multi-View Representation Learning.. In *ICML*. 1083–1092.

[42] Yishu Wang, Dejie Yang, and Minghua Deng. 2015. Low-rank and sparse matrix decomposition for genetic interaction data. *BioMed research international* 2015 (2015).

[43] World Health Organization. 2016. Dementia Fact sheet N362. Retrieved at https://web.archive.org/web/20150318030901/http://www.who.int/mediacentre/factsheets/fs362/en. (2016). Retrieved 13 January 2016.

[44] Tao Yang, Jun Liu, Pinghua Gong, Ruiwen Zhang, Xiaotong Shen, and Jieping Ye. 2016. Absolute Fused Lasso & Its Application to Genome-Wide Association Studies. In *Proceedings of the 22th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

[45] Tao Yang, Jie Wang, Qian Sun, Derrek P Hibar, Neda Jahanshad, Li Liu, Yalin Wang, Liang Zhan, Paul M Thompson, and Jieping Ye. 2015. Detecting genetic risk factors for Alzheimer's disease in whole genome sequence data via Lasso screening. In *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, 985–989.

[46] Lei Yuan, Yalin Wang, Paul M Thompson, Vaibhav A Narayan, Jieping Ye, Alzheimer's Disease Neuroimaging Initiative, and others. 2012. Multi-source feature learning for joint analysis of incomplete multiple heterogeneous neuroimaging data. *NeuroImage* 61, 3 (2012), 622–632.

[47] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Malsar: Multi-task learning via structural regularization. *Arizona State University* 21 (2011).

[48] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye. 2014. From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records. In *SIGKDD*. ACM, 135–144.

[49] Hongtu Zhu, Zakaria Khondker, Zhaohua Lu, and Joseph G. Ibrahim. 2014. Bayesian Generalized Low Rank Regression Models for Neuroimaging Phenotypes and Genetic Markers. *J. Amer. Statist. Assoc.* 109, 507 (2014), 977–990. DOI:http://dx.doi.org/10.1080/01621459.2014.923775