Longitudinal Self-Supervised Learning (LSSL)

Qingyu Zhao\textsuperscript{1*} \hspace{0.5cm} Zixuan Liu\textsuperscript{2*} \hspace{0.5cm} Ehsan Adeli\textsuperscript{1,3} \hspace{0.5cm} Kilian M. Pohl\textsuperscript{1,4}

\textsuperscript{1}Department of Psychiatry and Behavioral Sciences, Stanford University, CA 94305
\textsuperscript{2}Department of Electrical Engineering, Stanford University, CA 94305
\textsuperscript{3}Department of Computer Science, Stanford University, CA 94305
\textsuperscript{4}Center for Biomedical Sciences, SRI International, Menlo Park, CA 94205

\{qingyu, zucks626, eadeli, kilian.pohl\}@stanford.edu

Abstract

Longitudinal neuroimaging or biomedical studies often acquire multiple observations from each individual over time, which entails repeated measures with highly interdependent variables. In this paper, we discuss the implication of repeated measures design on unsupervised learning by showing its tight conceptual connection to self-supervised learning and factor disentanglement. Leveraging the ability for ‘self-comparison’ through repeated measures, we explicitly separate the definition of the factor space and the representation space enabling an exact disentanglement of time-related factors from the representations of the images. By formulating deterministic multivariate mapping functions between the two spaces, our model, named Longitudinal Self-Supervised Learning (LSSL), uses a standard autoencoding structure with a cosine loss to estimate the direction linked to the disentangled factor. We apply LSSL to two longitudinal neuroimaging studies to show its unique advantage in extracting the ‘brain-age’ information from the data and in revealing informative characteristics associated with neurodegenerative and neuropsychological disorders. For a downstream task of supervised diagnosis classification, the representations learned by LSSL permit faster convergence and higher (or similar) prediction accuracy compared to several other representation learning techniques.

1 Introduction

Longitudinal studies employ repeated measures to track a specific group of individuals over prolonged periods of time, often years or decades \cite{4}. When combined with neuroimaging techniques, these studies are particularly useful for evaluating the impact of age on the brain \cite{10}, the relationship between risk factors and development of disease \cite{20}, and the outcomes of treatments over time \cite{27}. Analyzing longitudinal data requires special computational tools, which have been traditionally grounded in statistical models, such as analysis of variance (ANOVA) \cite{5}. With recent advances in deep learning, supervised models, such as the Long Short-Term Memory (LSTM) networks \cite{19}, have become alternative approaches for analyzing longitudinal trajectory of individuals \cite{2} by formulating the problem as classification or prediction tasks.

Despite the enormous success of deep learning in longitudinal analysis, the implications of repeated measures design on unsupervised representation learning is unclear. In this paper, we aim to bridge

\*Equal Contribution

Preprint. Under review.
this gap by showing the tight conceptual connection from the repeated measures design to self-supervised learning and factor disentanglement. While the latter two topics are extensively studied in recent years, they are largely based on a cross-sectional design [39], where different training samples have no intrinsic relations with each other. In this situation, self-supervised learning [24] aims to automatically explore similarity and dissimilarity relations across samples through the learning process of their representations. This concept closely relates to the longitudinal design, in which each subject serves as his or her own ‘comparison’ with respect to the change over time. Therefore, learning representations from the repeated measures could fully leverage the structured inter-relation across time points.

Such an idea of learning time-dependent representations links to the concept of factor disentanglement. The key intuition behind disentanglement is that real-world data are generated by distinct, interpretable, informative underlying factors [41, 17]. A change in a single factor should lead to a change in a single dimension in the representation space. This notion, despite being appealing, cannot be stringently formulated for cross-sectional data. As pointed out in [28], disentanglement learning is theoretically unachievable unless with implicit supervision on the model. We argue that the repeated measures design is exactly proposed for ‘supervising’ the disentanglement of the evolving factor across the measures. In a longitudinal study where each individual is repeatedly examined over time, all time-independent factors are fixed so that the impact of time-related factors, such as age, can be effectively revealed on the observations.

Inspired by these notions, we propose a representation learning strategy, named Longitudinal Self-Supervised Learning (LSSL), to investigate the impact of aging and age-sensitive neuropsychological disorders on the brain in longitudinal neuroimaging studies. Thanks to the repeated measure design in longitudinal studies, we formulate disentanglement as a deterministic multivariate mapping from a ‘factor space’ to a representation space to explicitly model the impact of brain age on the representations. In practice, the only self-supervision of our model is to group images by subjects. Then, the training encourages the progression of subject-specific representations to follow a common developmental direction. This objective is condensed to optimizing a combination of image reconstruction loss and a simple cosine loss in the representation space. We test our model on two longitudinal MR data sets to investigate the impact of Alzheimer’s Disease (AD) and Alcohol Use Dependence (AUD) on the brain. While the training does not rely on the ground-truth age nor diagnosis labels, LSSL successfully disentangles the factor linked to brain age in the representation space and reveals accelerated aging effects of AD and AUD compared to the control cohort. When performing a downstream task of predicting diagnosis labels of subjects, the representations and pre-trained encoder learned by our model enable faster convergence and more accurate (or match up to) classification accuracy in a variety of settings compared to several commonly used state-of-the-art unsupervised or self-supervised representation learning strategies.

2 Related Works

Self-Supervised Learning for Images. Visual representation learning has recently focused on approaches utilizing self-supervised methods, which have demonstrated monumental gains compared to their supervised counterparts [7, 16, 24]. For instance, recent work has tackled representation learning in images using contextual approaches, such as exploiting color information [44, 25] and spatial relationships [40, 37]. Contrastive learning is proposed to explicit model similarity relationships across samples to regularize representation learning [7]. This idea is then extended to the modeling of temporal relationships [31]. Similar ideas have been used for super-resolution [12, 21] or estimating depth from stereo images [34]. Moreover, multi-task and cross-domain feature learning from imagery also benefits from training in a self-supervised manner [11, 35].

Longitudinal Neuroimaging Studies. Deep learning models applied to longitudinal neuroimaging studies are largely based on supervised learning, i.e., by training classification models based on image time series using regular recurrent networks [26, 38, 14, 8, 15]. Longitudinal pooling is proposed to augment the learned representation of each time point with information gathered from other images in the time series [32]. Methods for explicitly exploiting dependencies within the intra-subject series are mostly based on parameterizing the trajectories of representations in the latent space, such as using Mixed Effect Models [32, 29]. Unsupervised learning for longitudinal data involves deep clustering frameworks that form multi-modal parametric distributions in the representation space.
In general, unsupervised, weakly supervised, or non-parametric representation learning is a topic rarely studied for longitudinal neuroimaging data.

**Factor Disentanglement.** There is no consensus on the mathematical definition of ‘disentanglement’. Conceptually, a representation is considered disentangled if changes along one dimension in the representation space are explained by a specific factor of variation while being relatively invariant to other factors. Most exiting works formulate this notion from a statistical perspective by pursuing statistical independence among random variables in the latent space (factorizable latent representations [41, 17]). Therefore, state-of-the-art approaches for unsupervised disentanglement learning are based on a Variational Autoencoder (VAEs) structure, which aims to learn a factorizable posterior from the marginal distribution of the observed data [18, 6, 22, 48]. Despite the promising results from these works, the study by Locatello and colleagues [28] challenges the theoretical validity this the idea. They point out that given any marginal distribution of the observed data there exist an infinite number of generative processes from either disentangled or fully entangled latent representations. Therefore, factor disentanglement is theoretically impossible without having implicit supervision.

### 3 Theory

Current unsupervised learning frameworks often define the concept of ‘disentanglement’ as the observed data having factorizable (independent in each variable) distributions in the representation space, such that each representation variable corresponds to a real-world factor [18]. This learning objective is practically unrealistic because the assumption of statistical independence do not necessarily translate to factors observed in real world datasets. For example, brain morphology is influenced by both gender and brain size, two highly correlated factors. Pursuing factorizable distributions in the 2D representation space would yield two directions statistically independent (such as in Principal or Independent Component Analysis), but neither direction corresponds to real-world factors. To resolve this issue, we leverage the self-supervision enabled by the repeated measures to explicitly separate the concepts of a factor space and representation space. Here we assume images are generated by factors, and hence can be reduced to low-dimensional representations. Instead of imposing probabilistic assumptions on the two spaces, we define deterministic multivariate mapping functions as a means of disentanglement.

**Disentangled Mapping from Factor to Representation Space.** Let \( \Omega_\alpha = \mathbb{R}^M \) be the factor space, and \( \Omega_I = \mathbb{R}^P \) the image space. We assume each image \( I \in \Omega_I \) can be fully determined based on \( M \) factors \( \alpha = [\alpha_1, ..., \alpha_M] \in \Omega_\alpha \) through a differentiable generative process \( I = h(\alpha) \). Further, we aim to learn a differentiable encoder \( g \) that reduces the image to a low-dimensional representation \( z = [z_1, ..., z_K] \in \Omega_z \) in the representation space, where \( K \) is a given model parameter. We then define \( f : \Omega_\alpha \rightarrow \Omega_z \) as the composite multivariate-to-multivariate mapping \( f = g \circ h \), where each \( f_i \) is a differentiable multivariate-to-univariate mapping \( z_i = f_i(\tilde{\alpha}) \) for \( i \in [1, K] \).

Without loss of generality, we are interested in explicitly associating \( z_1 \) to the first factor \( \alpha_1 \). We consider \( \alpha_1 \) is disentangled from the representation if \( f \) can be factorized as

\[
z_1 = f_1(\alpha_1) \quad \text{and} \quad z_i = f_i(\tilde{\alpha}), \quad \text{for} \quad i > 1,
\]

where \( \tilde{\alpha} = [\alpha_2, ..., \alpha_M] \), and \( f_1 : \mathbb{R} \rightarrow \mathbb{R} \) is a strictly increasing function. In other words, disentanglement is achieved when 1) \( z_1 \) is solely dependent on \( \alpha_1 \) (the monotonicity of \( f_1 \) ensures

![Figure 1: Concept of our proposed Longitudinal Self-Supervised Learning (LSSL), which leverages the repeated measures to explicitly disentangle the factor of brain age in the representation space.](image)
the mapping to be bijective, i.e., without loss of information, and preserve ordinal information of factor \( z_1 \); and 2) the remaining representation in \( \mathbf{z} \) is solely dependent on factors other than \( \alpha_1 \).

**Self-Supervised Training.** In many applications, the only available data are the images \( \{ \mathbf{I} \} \). The underlying generative process, including the mapping function \( \mathbf{h} \), dimension of factor space \( \mathcal{M} \), and values of \( \alpha \) are hidden from observation, so training disentangled representation with respect to a specific factor becomes extremely challenging. However, in situations where each training sample has multiple images measured with respect to different values of a specific factor, we can leverage self-supervision to achieve disentanglement. To show this, the factorization of \( \mathbf{f} \) in Eq. (1) can be transformed to the following conditions:

\[
\frac{\partial f_i}{\partial \alpha_1} > 0, \quad \frac{\partial f_i}{\partial \alpha_1} = 0, \quad \frac{\partial f_i}{\partial \alpha_i} = 0 \text{ for } i > 1.
\]  

(2)

The above definition of factorization translates the problem setup on \( \mathbf{f} \) and \( \alpha \) to a setup with respect to \( \mathbf{g} \) and \( \mathbf{I} \). As

\[
\frac{\partial f_i}{\partial \alpha_j} = \sum_{p=1}^{P} \frac{\partial f_i}{\partial h_p} \frac{\partial h_p}{\partial \alpha_j} = \nabla \mathbf{u}_j g_i(\mathbf{I}), \quad \text{where } \mathbf{u}_j = [\frac{\partial h_1}{\partial \alpha_j}, ..., \frac{\partial h_P}{\partial \alpha_j}].
\]  

(3)

In other words, the partial derivative of a representation variable with respect to the change of a factor is equivalent to its directional derivative with respect to the corresponding change in the image space, where \( \mathbf{u}_j \) is the direction defined by the change of pixel (voxel) intensity values linked to the change in \( \alpha_j \). In practice, we let \( \mathbf{I}' \) be the resulting image after perturbing the value of \( \alpha_j \), then the corresponding change in \( z_i \) can be defined with respect to \( \mathbf{g} \) and \( \mathbf{I} \) with \( \mathbf{u}_j \approx \mathbf{I}' - \mathbf{I} \); i.e., \( \frac{\partial f_i}{\partial \alpha_j} \approx g_i(\mathbf{I}') - g_i(\mathbf{I}) \). As such, the disentanglement defined by Eq. (2) is achieved when

1. Upon perturbation of \( \alpha_1 \), \( \Delta \mathbf{z} = \mathbf{g}(\mathbf{I}') - \mathbf{g}(\mathbf{I}) \) is co-linear with \( [1, 0, ..., 0] \);
2. Upon perturbation of \( \alpha_i \) for \( i > 1 \), \( g_i(\mathbf{I}') - g_i(\mathbf{I}) = 0 \).

4 Longitudinal MR Studies for Brain Aging Analysis

Now we show how to apply the above strategy to longitudinal neuroimaging studies to inspect the impact of aging on the brain. To do so, we assume the phenotypic appearance of the human brain is determined by genotypic and demographic factors of individuals. Whern each individual is repeatedly scanned over time, the only factor that changes the brain morphology is the increase of ‘brain age’, while other commonly studied factors, such as gender and ethnicity, are fixed over time. Specifically, brain age characterizes the apparent health condition of the brain but not necessarily equals to chronological age (time since birth). For example, a patient with neurodegenerative disease can have a higher brain age than a healthy subject albeit they have the same chronological age.

**Optimization.** To disentangle the direction linked to brain age, we use a standard autoencoder structure to learn the encoder \( \mathbf{g} \), which requires learning an additional decoder \( \mathbf{d} \) to reconstruct the input image based on \( \mathbf{z} \), thus enforcing \( \mathbf{g} \) to learn informative representations beyond the single factor of interest. Further, we relax the assumption that the disentangled direction has to align with one natural coordinate axe in the representation space (as shown in [35], such alignment does not naturally result from the design of autoencoding) and parameterize the direction as a 1D unit vector \( \mathbf{\tau} \in \Omega_\tau \) that can be learned during training.

Let \( \mathbf{I}' \) and \( \mathbf{I}'' \) be two images from the same subject scanned at time \( s \) and \( t \) with \( s > t \). Let \( \psi'^t \) and \( \psi''^s \) be their projections to \( \mathbf{\tau} \), e.g., \( \psi'^t = \mathbf{g}(\mathbf{I}'^t; \theta) \downarrow \mathbf{\tau} = \mathbf{z}'^t \downarrow \mathbf{\tau} \). Then, ensuring Condition 1 while preserving \( \psi'' > \psi'^t \) is equivalent to enforcing \( \cos \langle \mathbf{g}(\mathbf{I}'^s; \theta) - \mathbf{g}(\mathbf{I}'^t; \theta), \mathbf{\tau} \rangle = 1 \), i.e., a zero-angle between \( \mathbf{\tau} \) and the direction of progression in the representation space. To impose this constraint in the autoencoder, we propose to add the cosine loss for each image pair to the objective function

\[
\min_{\theta, \phi} \sum_{\mathbf{I} \in \mathcal{I}} \text{MSE} \left( \mathbf{I}, \mathbf{d} \left( \mathbf{g} \left( \mathbf{I}^t; \theta \right) ; \phi \right) \right) - \lambda \sum_{\mathbf{I}', \mathbf{I}'' \in \mathcal{S}} \cos \langle \mathbf{g}(\mathbf{I}'^s; \theta) - \mathbf{g}(\mathbf{I}'^t; \theta), \mathbf{\tau} \rangle,
\]  

(4)

where \( \theta \) and \( \phi \) are the parameters of the neural networks, \( \mathcal{I} \) is the image repository, \( \mathcal{S} \) is the set of subject-specific image pairs, and \( \lambda \) is the parameter weighting the two terms. As a result, the cosine
loss encourages colinearity among the development of brain representation across subjects while assigning directionality to $\tau$ (higher $\psi$ indicates older brain age).

Note, optimizing Eq. (4) only leads to a necessary condition (Condition 1) for disentanglement defined in Section 3. This is because the longitudinal design only considers a single varying factor across the repeated measures and therefore does not provide information for Condition 2. However, when factor $\alpha_i$ is known to be statistically independent of $\alpha_j$ (such as gender vs. brain age), $\alpha_i$ is also independent of the disentangled coordinates $\psi$ associated with $\alpha_j$. We can use this criterion to validate the disentanglement of brain age with respect to other demographic factors.

5 Experiments

We evaluate the proposed model on two longitudinal neuroimaging datasets: the Alzheimer’s Disease Neuroimaging Initiative (ADNI1) and a data set on Alcohol Use Dependence (AUD). All longitudinal MRIs in the following experiments were first preprocessed by a pipeline composed of denoising, bias field correction, skull striping, affine registration to a template, re-scaling to a $64 \times 64 \times 64$ volume, transforming image intensities within the brainmask to z-scores [46]. We design an encoder composed of 4 stacks of $3 \times 3 \times 3$ convolution/ReLU/max-pooling layers with dimension (16,32,64,16). Then, a fully connected layer results in a 512 dimensional representation space. The decoder employs a reverse structure of the encoder.

Evaluation. We first correlate the disentangled brain age with ground-truth chronological age and show how such disentanglement can reveal characteristics with respect to the specific brain disorder under investigation. This analysis is unique to LSSL and cannot be achieved by the baselines. We then evaluate the quality of the learned representations by using them for predicting the diagnosis label of individuals. The classification is evaluated by 5-fold cross-validation in both cross-sectional (i.e., based on single time point) and longitudinal settings. The 5 folds are split based on subjects; i.e., images of a single subject belong to the same fold. After splitting folds, the cross-sectional model discards the temporal information within each subject and treats each image as an independent sample. The classification model is a Multi-Layer Perceptron containing two fully connected layers of dimension 512 and 64 with ReLU activation. On the other side, the longitudinal model is a RNN that trains and predicts based on the sequence of images for each subject. The RNN maps the latent representation to a 16D vector, which is fed into a single layer GRU network with 16 hidden units. Finally, we re-evaluate the two classification settings in an end-to-end manner by fine-tuning the representations (encoder) pre-trained by LSSL.

Baselines. We compare the classification accuracy and the converging speed in the fine-tuning setting with encoders pre-trained by several other state-of-the-art representation learning methods. As LSSL is conceptually related to a wide range of works, we select several representative methods in unsupervised training (AE and VAE), factor disentanglement ($\beta$-VAE [18]), self-supervised learning (SimCLR [7]), and a longitudinal framework based on Contrastive Predictive Coding (CPC [31]). Note that the pre-training of CPC already contains an auto-regressive model on top of the encoder, which reduces the representation to a 16D vector followed by a GRU with 16 hidden units, known as the context features. Therefore, the longitudinal prediction of CPC is directly based on the context features instead of the 512D representations.

5.1 Longitudinal Study of Alzheimer’s Disease

We first evaluate the proposed model on the longitudinal structural MRIs from ADNI1, which consists of 229 normal control subjects (age: $76 \pm 5.0$ years), 397 subjects diagnosed with Mild Cognitive Impairment ($74.9 \pm 7.4$ years), and 185 subjects with Alzheimer’s Disease ($75.3 \pm 7.6$ years). Each subject has 1 to 8 longitudinal scans within a 4 year study period. To train our model, we select all image pairs with each belonging to the same subject and having at least 1 year interval between the scanning time. The value of $\lambda = \frac{|S|}{|I|}$ (Eq. 4) so that there is a balance between the number of images ($|I|$) and image pairs ($|S|$) in the training. Note the self-supervised training only considers the temporal order of each image pair, but did not use the ground-truth diagnosis label or the chronological age of the subjects.

that successfully passed through our image-preprocessing pipeline
Table 1: Average cross-sectional and longitudinal classification accuracy with and w/o fine-tuning the encoder. Best result in each column has bold typeset, and the second best is underlined.

| Pre-training Model | ADNI  | AUD  |
|--------------------|-------|------|
|                    | CNN   | CNN+RNN | CNN   | CNN+RNN |
| AE                 | 58.6 | 81.7 | 62.1 | 71.3 | 58.8 | 69.1 | 52.1 | 53.2 |
| β-VAE [18]        | 58.9 | 75.7 | 62.8 | 71.9 | 55.4 | 70.2 | 63.0 | 65.6 |
| SimCLR [7]        | 56.1 | 77.2 | 76.3 | 78.4 | 52.1 | 67.5 | 60.8 | 61.0 |
| VAE [23]          | 58.9 | 75.7 | 62.8 | 71.9 | 55.4 | 70.2 | 63.0 | 65.6 |
| β-VAE [18]        | 56.1 | 77.2 | 76.3 | 78.4 | 52.1 | 67.5 | 60.8 | 61.0 |
| SimCLR [7]        | 58.9 | 75.7 | 62.8 | 71.9 | 55.4 | 70.2 | 63.0 | 65.6 |
| Ours (LSSL)       | 72.0 | 84.1 | 81.8 | 87.0 | 62.9 | 71.7 | 67.0 | 72.0 |

We first test the model by computing the brain age $\psi$ of the 2,641 images from the 811 subjects. As the parameterization of the disentangled direction $\tau$ is uniquely determined up to scale, we normalize $\psi$ such that its mean and standard deviation is matched to those of chronological age in the dataset. Fig. 2a shows the brain age of the control subjects versus their chronological age. According to the fitting of a quadratic mixed effect model, brain age and chronological age exhibit a nearly linear relationship over the entire age span of the dataset, indicating that the factor linked to brain age is successfully disentangled from the learned representation. Note this global relationship is learned solely from the ordinal information from image pairs (maximum 4 years part) without using the ground-truth age of subjects as commonly done in supervised learning [45, 9]. As mentioned, brain age is not necessarily equivalent to chronological age but is solely a marker of health condition of the brain. This is indicated by variance of brain age for any given chronological age in Fig. 2a.

According to Fig. 2b, the brain age of AD patients is generally higher than chronological age reflecting the neurodegenerative nature of AD that accelerates brain aging. This phenomenon can also be inferred from Fig. 2c, where we compute a slope measure by fitting simple linear regression for each subject (with at least 2 images) on their brain age across visits. In doing so, we see that the control group has an average aging speed (slope) close to 1, indicating the consistency between the progression rate of brain age and of chronological age. On the other hand, AD group has significantly faster brain aging than normal ($p < 0.001$, two-sample t-test). Interestingly, the MCI group, representing a transitional state between control and AD, has an intermediate aging speed, which is significantly faster than normal ($p < 0.001$) and slower than AD ($p < 0.001$). Lastly, we observe that the gap between brain age and chronological age is larger in younger AD patients than the older ones (Fig. 2b) indicating the younger brain is more vulnerable to the disease. This effect is quantitatively supported (Fig. 2d) by fitting a linear regression between aging speed and age, which reveals that the AD group has accelerated aging in the younger cohort but converge with normal in the later stage. Again, the MCI group shows an intermediate effect between AD and normal.

We also assess the quality of disentanglement by simulating the average brain at different brain ages. To generate an average representation for brain age $\psi$, we construct the age-related component by sampling along $\tau$ and construct the age-independent component as the group average of the

Figure 2: Brain age of 229 control subjects from ADNI1 (a) and of 185 AD patients (red) overlaid with normal developmental trajectory (b). Speed of brain aging (slope of $\psi$ over time) for the 3 diagnosis groups (c) and as a function of chronological age of the ADNI1 subjects (d).
components orthogonal to $\tau$. $z^{\psi} = \psi + \frac{1}{N} \sum_{i=1}^{N} (z^i - z^i \tau^\top \tau)$. By decoding the age-dependent representation, we observe a pattern of enlargement in the ventricle and loss of brain tissues as age increases (Fig. 4), which converges with the concept in current neuroscience literature [40].

Lastly, we classify AD patients from control subjects based on the learned representations. For fair comparison, we use a fixed network architecture for all methods in this experiment, as our goal is to show the superiority of our self-supervised representation learning rather than obtaining state-of-the-art results on ADNI dataset with complete architecture design. In the cross-sectional setting, the representations learned by LSSL enable more accurate prediction than the baselines except for SimCLR (Table 1 CNN frozen). This is to be expected because the CNN classification neglects temporal correlation across images of each individual, so our longitudinal modeling might not show full advantage. Note that the accuracy of our model closely matches up to SimCLR after fine-tuning the encoder in the cross-sectional setting. On the other hand, when performing classification based on the longitudinal sequences (CNN+RNN), the accuracy associated with the LSSL representations significantly increases and outperforms the baselines, which is also the case after fine-tuning the encoders. This indicates that the representations learned by LSSL contain more informative temporal relations across subject-specific images. Moreover, when performing the fine-tuning in an end-to-end setting, the encoder pre-trained by LSSL demonstrates faster convergence rate in both cross-sectional and longitudinal settings despite that all methods start from random predictions (due to random initialization of the MLP layers) (Fig. 5).

5.2 Longitudinal Study of Alcohol Use Disorder

Another neurological disorder known to accelerate brain aging is Alcohol Use Disorder (AUD), which can cause gradual deterioration in the gray and white matter tissue [33, 43]. In this experiment, we apply the proposed model to the longitudinal T1-weighted MRIs (up to 13 visits) of 274 Normal Controls (age: $47.3 \pm 17.6$) and 329 patients diagnosed with AUD (age: $49.3 \pm 10.5$).

Similar to the previous experiment, we first visualize the brain age of the control subjects. Compared to the linear aging pattern between age 60 to 90 years revealed in the ADNI control subjects, the aging of the AUD control subjects exhibits a quadratic pattern between age 20 to 85.
years, where younger adults age slower compared to the older subjects (Fig. 5a). To put the results in context, we normalize the projection $\psi$ with respect to the age range of the ADNI dataset. Fig. 5a shows coherent aging trajectories between the two datasets despite the distinction in the age span of cohorts and imaging protocols. Similar to the ADNI experiment, AUD patients also exhibit higher brain age than normal controls (Fig. 5b) supported by their slopes (aging speed) significantly larger than normal ($p < 0.001$, Fig. 5c). Different from the ADNI experiment is that the aging speed of AUD patients is always larger than normal controls regardless of their chronological age (Fig. 5d). This results in an accumulative alcohol effect (Fig. 5b), where older subjects have a larger gap between brain and chronological age. These results comply with the fact that chronic drinking often gradually deteriorates brain structure resulting in more severe alcohol effect in older subjects [47].

In Fig. 4, we simulate images of different brain ages for the AUD dataset. We observe that the simulated brains closely mimics the ones from the ADNI experiment from age 60 to 70, a range where the two datasets highly overlap. However, the simulated brains from the AUD experiment show less pronounced aging effect after age 70 compared to the ADNI results. This is potentially due to the few older subjects in the AUD dataset compared with ADNI, so the model conservatively extrapolates the aging pattern for the older age range.

Prior literature indicates that AUD is only weakly separable from the control group [1, 32], which is echoed in our results by the significantly lower accuracy in the frozen setting (all methods < 65%, Table 1) compared to the ADNI experiment. However, LSSL still has the fastest converging rate and highest accuracy upon convergence in the fine-tuning setting compared to the baselines (Fig. 6). As indicated in our results and a prior study by Alberti and colleagues [3], fully unsupervised pre-training (e.g., VAE-based) does not necessarily guarantee better performance in supervised downstream tasks. The challenging AUD classification also takes the RNN longer to converge for several baselines, which is not the case for LSSL.
6 Conclusion

In this work, we proposed a self-supervised representation learning framework called LSSL that incorporated theoretical advantages from the repeated measures design in longitudinal neuroimaging studies. The explicit longitudinal self-supervision permitted separate definitions for the factor and representation spaces, thereby omitting the ambiguity often encountered in fully unsupervised disentanglement models. Based on optimizing the colinearity between a global direction in the representation space and the developmental trajectories from subject-specific image pairs, LSSL successfully disentangled the factor of brain aging in the representation space, which was used to characterize normal aging pattern across the life span and to reveal the accelerated aging effects of Alzheimer’s Disease and Alcohol Use Dependence. Compared to several other state-of-the-art representation learning methods, the pre-trained encoder and representations learned by LSSL are more suitable for supervised classification of diagnosis labels in various settings, indicated by faster convergence and higher (or equally high) prediction accuracy upon convergence.

Broader Impact

Unsupervised learnable representations from neuroimages can play a crucial role in accelerating neuroscientific discovery by avoiding a priori assumptions. Factor disentanglement, when coupled with self-supervised learning in the longitudinal setting, can provide interpretable representations in a fully data-driven fashion, which is an essential component in neuroscience. Two of the main reasons that prevented deep learning and in general complex data-driven methods to be applied for neuroscience discoveries were (1) hardship in interpreting the obtained models and (2) the need for large sets of labelled data for training model (for which medical studies in general suffer due to patient privacy concerns and limited data sharing policies). Methods like ours can play an important role towards learning interpretable representations leveraging longitudinal self-supervision, which also has the potential to shed light on general machine learning tasks that require temporal modeling, such as gait analysis.

References

[1] Ehsan Adeli, Dongjin Kwon, and Kilian Pohl. Multi-label Transduction for Identifying Disease Comorbidity Patterns: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part III, pages 575–583. 09 2018.
[2] Maryamossadat Aghili, Solale Tabarestani, Malek Adjouadi, and Ehsan Adeli. Predictive modeling of longitudinal data for alzheimer’s disease diagnosis using rnns. In International Workshop on PRedictive Intelligence In MEdicine, pages 112–119. Springer, 2018.
[3] Michele Alberti, Mathias Seuret, Rolf Ingold, and Marcus Liwicki. A pitfall of unsupervised pre-training, 2017.
[4] Edward Caruana, M. Roman, Jules Hernández-Sánchez, and Piergiorgio Solli. Longitudinal studies. Journal of thoracic disease, 7:E537–40, 12 2015.
[5] Philip Cash, Tino Stanković, and Mario vStorga. An Introduction to Experimental Design Research, pages 3–12. 06 2016.
[6] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2018.
[7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.
[8] Ruoxuan Cui, Manhua Liu, Alzheimer’s Disease Neuroimaging Initiative, et al. Rnn-based longitudinal analysis for diagnosis of alzheimer’s disease. Computerized Medical Imaging and Graphics, 73:1–10, 2019.
[9] Adrian V . Dalca, Marianne Rakic, John Guttag, and Mert R. Sabuncu. Learning conditional deformable templates with convolutional networks, 2018.
[10] Nancy Dennis and Roberto Cabeza. Neuroimaging of healthy cognitive aging. Fundamenta Informaticae - FUIN, 3, 01 2008.
[11] Carl Doersch and Andrew Zisserman. Multi-task self-supervised visual learning. In Proceedings of the IEEE International Conference on Computer Vision, pages 2051–2060, 2017.
Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European conference on computer vision*, pages 184–199. Springer, 2014.

Louis Falissard, Guy Fagherazzi, Newton Howard, and Bruno Falissard. Deep clustering of longitudinal data, 2018.

Ruiqiang Gao, Yuankai Huo, Shunxing Bao, Yucheng Tang, Sanja L. Antic, Emily S Epstein, Aneri B Balar, Steve Deppen, Alexis B Paulson, Kim L Sandler, et al. Distanced lstm: Time-distanced gates in long short-term memory models for lung cancer detection. In *International Workshop on Machine Learning in Medical Imaging*, pages 310–318. Springer, 2019.

Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, M Jorge Cardoso, Marc Modat, Sébastien Ourselin, Lauge Sørensen. Alzheimer’s Disease NeuroImaging Initiative, et al. Training recurrent neural networks robust to incomplete data: Application to alzheimer’s disease progression modeling. *Medical image analysis*, 53:39–46, 2019.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.

Irina Higgins, David Amos, David Pfau, Sébastien Racaniere, Loïc Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations, 2018.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Clifford Jack, Matt Bernstein, Nick Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula Britson, Jennifer Whitwell, Chadwick Ward, Anders Dale, Joel Felmlee, Jeffrey Gunter, Derek Hill, Ron Killiany, Norbert Schuff, Sabrina Fox-Bosetti, Chen Lin, Colind Studholme, and Michael Weiner. The alzheimer’s disease neuroimaging initiative (adni): Mri methods. *Journal of magnetic resonance imaging: JMRI*, 27:685–91, 05 2008.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

A. Kolesnikov, X. Zhai, and L. Beyer. Revisiting self-supervised visual representation learning. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1920–1929, 2019.

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. Colorization as a proxy task for visual understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6874–6883, 2017.

Zachary C Lipton, David C Kale, Charles Elkan, and Randall Wetzel. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677*, 2015.

Chunyan Liu, Tim Cripe, and Mi-Ok Kim. Statistical issues in longitudinal data analysis for treatment efficacy studies in the biomedical sciences. *Molecular therapy: the journal of the American Society of Gene Therapy*, 18:1724–30, 09 2010.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2018.

Maxime Louis, Raphael Couronné, Igor Koval, Benjamin Charlier, and Stanley Durrleman. Riemannian geometry learning for disease progression modelling. In *Information Processing in Medical Imaging*, pages 542–553, 05 2019.

Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European Conference on Computer Vision*, pages 69–84. Springer, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Jiahong Ouyang, Qingyu Zhao, Edith V Sullivan, Adolf Pfefferbaum, Susan F. Tapert, Ehsan Adeli, and Kilian M Pohl. Recurrent neural networks with longitudinal pooling and consistency regularization, 2020.

Adolf Pfefferbaum, Margaret Rosenbloom, Weiwei Chu, Stephanie Sassoon, Torsten Rohlfing, Kilian Pohl, Natalie Zahr, and Edith Sullivan. White matter microstructural recovery with abstinence and decline with relapse in alcohol dependence interacts with normal ageing: A controlled longitudinal dti study. *The Lancet Psychiatry*, 1:202–212, 08 2014.
[34] Sudeep Pillai, Rarees Ambrues, and Adrien Gaidon. Superdepth: Self-supervised, super-resolved monocular depth estimation. In 2019 International Conference on Robotics and Automation (ICRA), pages 9250–9256. IEEE, 2019.

[35] Zhongzheng Ren and Yong Jae Lee. Cross-domain self-supervised multi-task feature learning using synthetic imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 762–771, 2018.

[36] Michal Rolinek, Dominik Zietlow, and Georg Martius. Variational autoencoders pursue pca directions (by accident), 2018.

[37] Mohammad Sabokrou, Mohammad Khalooei, and Ehsan Adeli. Self-supervised representation learning via neighborhood-relational encoding. In Proceedings of the IEEE International Conference on Computer Vision, pages 8010–8019, 2019.

[38] Ruggiero Santeramo, Samuel Withey, and Giovanni Montana. Longitudinal detection of radiological abnormalities with time-modulated lstm. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 326–333. Springer, 2018.

[39] Maninder Setia. Methodology series module 3: Cross-sectional studies. Indian Journal of Dermatology, 61:261–264, 05 2016.

[40] Edith Sullivan and A Pfefferbaum. Neuroradiological characterization of normal adult aging. The British journal of radiology, 80 Spec No 2:S99–108, 01 2008.

[41] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. 12 2018.

[42] Yunyang Xiong, Hyunwoo Kim, Ronak Mehta, Sterling Johnson, and Vikas Singh. On training deep 3d cnn models with dependent samples in neuroimaging. In Information Processing in Medical Imaging, pages 99–111, 05 2019.

[43] Natalie M. Zahr and Adolf Pfefferbaum. Alcohol’s effects on the brain: Neuroimaging results in humans and animal models. Alcohol Res., 38(2):183–206, 2018.

[44] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In European conference on computer vision, pages 649–666. Springer, 2016.

[45] Qingyu Zhao, Ehsan Adeli, Nicolas Honnorat, Tuo Leng, and Kilian M. Pohl. Variational autoencoder for regression: Application to brain aging analysis. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 823–831, 2019.

[46] Qingyu Zhao, Ehsan Adeli, Adolf Pfefferbaum, Edith V. Sullivan, and Kilian M. Pohl. Confounder-aware visualization of convnets. In International Workshop on Machine Learning in Medical Imaging, pages 328–336, 2019.

[47] Qingyu Zhao, Adolf Pfefferbaum, Simon Podhajsky, Kilian M. Pohl, and Edith V. Sullivan. Accelerated aging and motor control deficits are related to regional deformation of central cerebellar white matter in alcohol use disorder. Addiction Biology, 25(3), 04 2019.

[48] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Information maximizing variational autoencoders, 2017.