Predictive Image Regression for Longitudinal Studies with Missing Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this paper, we present a predictive regression model for longitudinal images with missing data based on large deformation diffeomorphic metric mapping (LDDMM) and deep recurrent neural network. Instead of directly predicting image scans, our model predicts a vector momentum sequence of an baseline image, because it parameterizes the corresponding image sequence and lies in the tangent space of the baseline images, which is Euclidean. A neural network with long term-short memory (LSTM) units is applied to learn the time-varying changes from the vector-momentum sequences, which are generated by using LDDMM. For the baseline image that the vector momenta are associated with, it is encoded by a convolutional neural network (CNN). Both features from the LSTM and CNN are fed into a decoder network to reconstruct the vector momentum sequence, which will be used to deform the baseline image and generate the corresponding image sequence with LDDMM shooting. To handle the missing images at some time points, a mask is adopted to ignore their reconstructions. We show our results on synthetically generated images and the brain MRIs from the OASIS dataset. Our method accurately predicts the spatio-temporal changes in both datasets, irrespective of large or subtle changes in longitudinal image sequences.

1 Introduction

Since the last decade longitudinal images are increasingly available for studying brain development and degeneration, disease progression, and aging problems. For instance, to understand the time-varying evolution of longitduinal data like human brains over years, image regression [1] is a commonly-used technique to capture underlying spatial-temporal changes. This regression model estimates images as a function of associated variables like age under the framework of Large Deformation Diffeomorphic Metric Mapping (LDDMM) [2]. The following-up works aim at capturing non-linear changes with polynomial or spline regression [3, 4], modeling hierarchical changes at subject- and group-levels separately [5], or improving computational efficiency by introducing model approximations [6, 7]. These methods summarize the time-varying changes of a population, which is parameterized by the initial conditions of the captured smooth trajectory, e.g., the initial image and its associated initial momentum or velocity. To leverage this summarized trajectory and predict follow-up image scans for a specific subject, parallel transport techniques like [8] are required to transport the estimated initial momentum from its initial image to the image of that subject; however, this is non-trivial and the techniques are still under development. Image regression approaches based on kernel regression have also been explored [9, 10], but they cannot provide an explicit model for further statistical analysis and the prediction depends on the whole training data which is not efficient.

Recent advances and success in deep neural networks (DNN) [11] provide an alternative strategy to study longitudinal image populations. Several convolutional neural networks (CNN) and recurrent

Submitted to 1st Conference on Medical Imaging with Deep Learning (MIDL 2018). Do not distribute.

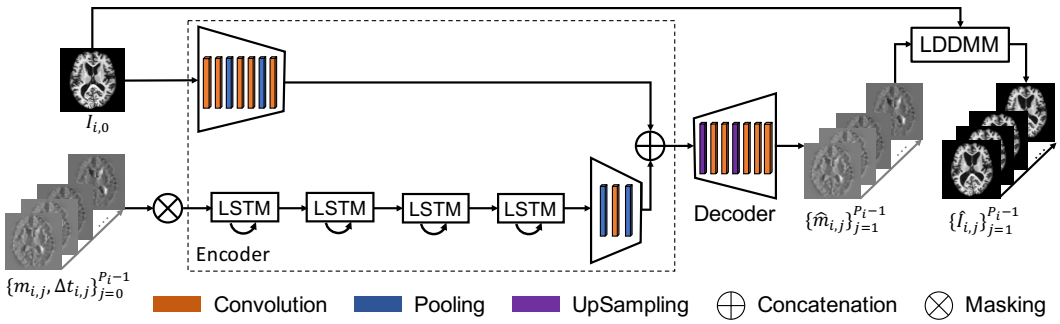

Figure 1: Architecture of our predictive image regression network. The baseline image $I_{i,0}$ of a subject i passes through a CNN image encoder to extract its features, which are concatenated with features of a sequence of vector momenta $\{m_{i,j}\}$ through an LSTM encoder. The 0/1 mask is used to ignore missing data at some time points $\{t_{i,j}\}$. The concatenated feature maps are the input of a CNN decoder to reconstruct the vector momentum sequence $\{\hat{m}_{i,j}\}$, which is used to deform the image $\{I_{i,0}\}$ and predict the image sequence $\{\hat{I}_{i,j}\}$ using LDDMM.

neural networks (RNN) [12, 13] have been proposed to predict the next frame of a video, without requiring the estimation of complex mathematical regression models. Since they treat images as a collection of pixel intensities without understanding their geometrical structures, these data-driven methods based on DNNs often suffer from the prediction of blurry images. Furthermore, different from video frame prediction, our task of predicting medical image scans tackles longitudinal data collected at different time points with varying time intervals, in particular, subjects have scans at different ages and each subject may not be scanned every year. That is, we have missing data issue, which does not often exist in video prediction.

In this paper, we consider the problem of predicting follow-up image scans of a specific subject with a baseline scan as input, using a neural network learned from longitudinal data with missing scans. To overcome the blurry prediction issue of DNNs, we integrate image registration techniques with a mixed CNN and RNN architecture (as shown in Fig. 1). Instead of directly predicting image sequence of a subject, our model predicts a sequence of vector momenta [14] for a given baseline image ($\mathcal{I}_0$). These vector momenta are associated with the baseline image and are used to deform it to different time points and generate the corresponding image sequence, under the LDDMM framework. The network is trained by the first image of a subject and its associated vector momentum sequence for the follow-up scans, which are generated using LDDMM image registration. To handle the missing data, a 0/1 mask is introduced during the training procedure. For an individual with one image scan, our model can predict how it changes over time.

The most related work to ours is the fast image regression in [15], which uses pairwise fast image registrations [16] to estimate a simplified image regression model [7]. The model proposed in [7] uses distance approximation for measuring image differences based on the assumption of only small deformations existing among images. To relax this assumption, our model fully leverages longitudinal data and learns the deformations from time series data. This is achieved by adopting an RNN composed of Long Short Term Memory (LSTM) units, which is well suited for time-series prediction problems. Our model is evaluated on both synthetic and real datasets and the results demonstrate the effectiveness of our proposed model.

2 Predictive Network for Image Regression

Assume we have a population of images collected from N subjects and each subject i has a varying number (P_i) of images ($\{I_{i,j}\}_{j=0}^{P_i-1}$) scanned at different time points ($\{t_{i,j}\}_{j=0}^{P_i-1}$). The objective of image regression is to uncover the relationship between images $\{I_{i,j}\}$ and their associated variable $\{t_{i,j}\}$. Instead of directly working on images, we leverage LDDMM and geodesic shooting with vector momentum [17, 14] to convert the longitudinal images of a subject into an initial image $I_{i,0}$ and a sequence of associated momenta $\{m_{i,j}\}_{j=0}^{P_i-1}$ (Section 2.1). The relation of the initial image

and its momentum sequence to their associated variables, like age, will be learned through training a neural network (Section 2.2). The predicted momentum sequence for an input image can shoot it forward to generate the corresponding image sequence.

2.1 LDDMM and Momentum Generation

To study image time series, we should first establish the correspondences between images and the LDDMM [2] framework provides a solution with maps of diffeomorphisms (smooth mapping and smooth inverse mapping) to deform one image to another. Specifically, given a source image $\mathcal{I}_0$ and a target image $\mathcal{I}_1$, the LDDMM formulation aims to minimize the following energy:

$$E(v) = \int_0^1 \|v\|_L^2 dt + \frac{1}{\sigma^2}\|\mathcal{I}_0 \circ \Phi^{-1}(1) - \mathcal{I}_1\|_2^2,$$
$$s.t. \quad \frac{d\Phi}{dt} = v \circ \Phi,\ \Phi^{-1}(0) = Id.$$

Here, v is a spatiotemporal velocity field, L is a differential operator on the velocity field to enforce its smoothness, e.g., $L = -\alpha\nabla^2 - \beta\nabla(\nabla\cdot) + \gamma$, $\sigma > 0$ is a constant to balance the first regularization term and the second image matching term in the above equation, Φ is the diffeomorphic mapping, and Id is the identity map. This formulation can be solved using the shooting strategy [17] and the image registration from $\mathcal{I}_0$ to $\mathcal{I}_1$ can be parametrized by a initial vector momentum m_0. The m_0 is associated with $\mathcal{I}_0$ and is the dual of the velocity field, that is, $m = Lv$, as proposed in [14].

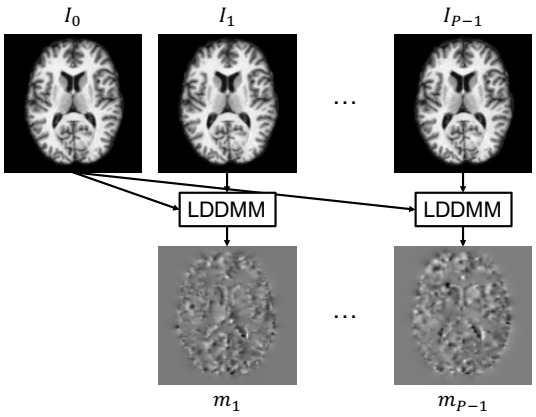

Figure 2: Generation of a vector momentum sequence using LDDMM. The baseline image ($\mathcal{I}_0$) is paired with every other image in the sequence ($\mathcal{I}_1$, $\mathcal{I}_2$, ..., $\mathcal{I}_{P-1}$).

Given a sequence of longitudinal images from a subject i, we select the first scan $I_{i,0}$ as the baseline image and compute the initial vector $m_{i,j}$ for each scan $I_{i,j}$, $(j = 0 \cdots P_i - 1)$, by registering $I_{i,0}$ to $I_{i,j}$ using LDDMM, as shown in Fig.2. Each vector momentum has x, y, and z components and the z dimension has zero momentum for a 2D image. As a result, we represent the image sequence of a subject as one image and its associate vector momentum sequence. The first momentum is registering the baseline image to itself, which is zero and serves as a starting point for our recurrent network prediction discussed in Section 2.2. Each vector momentum inherits the associated variable of its corresponding image. That is, we have a population of data represented as $\{I_{i,0}, \{m_{i,j}, t_{i,j}\}_{j=0}^{P_i-1}\}_{i=0}^{N-1}$ by using LDDMM and geodesic shooting.

2.2 Predictive Regression Network

The new data representation we obtain from Section 2.1 brings longitudinal images of a subject, which are deformed on the manifold of diffeomorphisms, to the tangent space of the first scan which has Euclidean structure. That is, the vector momenta are Euclidean data and its relationship to the associated variable can be learned using an RNN, which is designed to handle time series data. At the same time, the vector momenta are associated with the first image scan, which should also be included for understanding the momentum sequence. Therefore, we design an encoder-and-decoder network for solving our image regression task, as shown in Fig. 1. The encoder has a CNN to extract the baseline image features and an RNN to handle the vector momentum sequences. The decoder reconstructs the vector momenta, which will be used to shoot the baseline image forward to generate follow-up image scans.

CNN Image Encoder. The feature extractor is a series of convolution and pooling layers to learn features from the baseline image $I_{i,0}$ of a subject i. The initial pair of convolution layers have 32 filters and the second pair increase to 64. A max pooling layer and dropout layer follow every pair of

Algorithm 1 Workflow of Our Predictive Network for Image Regression

1: **procedure** PREPROCESSING
2: Compute the vector momentum sequence $m_0, m_1, ..., m_{P_i-1}$ of each subject i using LDDMM for image pairs $\mathcal{I}_0 \rightarrow \mathcal{I}_0, \mathcal{I}_0 \rightarrow \mathcal{I}_1, \mathcal{I}_0 \rightarrow \mathcal{I}_2$, ..., $\mathcal{I}_0 \rightarrow \mathcal{I}_{P_i-1}$, as shown in Fig. 2.
3: Construct vector momentum sequences such that m_0 has a target m_1, m_0 and m_1 together have a target sequence m_2, and so on.
4: Save the initial images $\mathcal{I}_0$ with those vector momentum sequences.

5: **procedure** TRAINING
6: Pre-train the CNN vector momentum decoder using computed vector momenta from step 2. This is achieved by training an autoencoder to reconstruct the vector momenta.
7: Train the LSTM encoder on vector momentum sequences constructed from step 3.
8: Train the CNN image encoder on initial images ($\mathcal{I}_0$) of all subjects through training an autoencoder to reconstruct the images.
9: Merge the features extracted from steps 7 and 8 and feed them to the decoder.
10: Fine-tune the weights of the decoder by training it over the merged features from the two encoders.

11: **procedure** PREDICTION
12: Extract features from input image $\mathcal{I}_0$ using the image encoder network.
13: Set m_0 to be zero.
14: Feed m_0 to the LSTM encoder network.
15: Feed extracted features from the input image and initial momentum m_0 to the decoder.
16: The decoder predicts the next vector momentum $\hat{m}_1$.
17: Append $\hat{m}_1$ and form a momentum sequence with m_0.
18: Extract features from this newly formed sequence by giving it as an input to the LSTM network to predict $\hat{m}_2$.
19: Repeat from step 14 to step 18 to predict $\hat{m}_3$, $\hat{m}_4$, ..., until $\hat{m}_{s-1}$ is predicted. Here, s is the maximum sequence length that the LSTM network can handle.
20: Apply the predicted vector momentum sequence to $\mathcal{I}_0$ with LDDMM shooting to generate the sequence of the follow-up images $\hat{\mathcal{I}}_1$, $\hat{\mathcal{I}}_2$, ..., $\hat{\mathcal{I}}_{s-1}$.
21: If needed, treat $\hat{\mathcal{I}}_{s-1}$ as $\mathcal{I}_0$ and $\hat{m}_{s-1}$ as m_0 to continue the prediction.

convolution units. The last convolution layer in this CNN has 128 filters. Throughout the network, we use PReLU activation function [18], the convolution units with 3×3 kernel size, and a pooling size of 2×2. We choose PReLU activation function because of negative values in the vector momenta. At the end, the extracted features are concatenated with the output from the LSTM vector momentum encoder and fed to the CNN vector momentum decoder for generating momentum sequence.

LSTM Vector Momentum Encoder. The vector momentum sequences $\{m_{i,j}, t_{i,j}\}$, which are generated by LDDMM and geodesic shooting, are fed to a recurrent network with LSTM units for learning the vector momentum as a function of time. Since the changes start from the baseline image at $t_{i,0}$, we adjust this independent variable as the age difference in years to the baseline image, that is, $\Delta t_{i,j} = t_{i,j} - t_{i,0}$. The maximum age difference accepted by the LSTM units is five years. That is, we consider longitudinal images collected within five years. This number is determined by the OASIS dataset (see Section 3) and can be adjusted to a new longitudinal dataset.

In this design, the age difference between adjacent LSTM units is one year. In practice, it is very likely that a subject misses a scan in some following-up years, which happens in the OASIS dataset. To deal with the missing data, a masking layer is added after the input layer to mask the missing inputs in the sequence. The LSTM vector momentum encoder is similar to the CNN image encoder in terms of the composition of convolution layers. In this recurrent network, a total of four convolutional LSTM layers follow the masking layer. Among the four convolutional LSTM units, the first two units have 32 filters, which is then doubled to 64 for the next pair. A Batch Normalization layer is added after every convolutional LSTM layer. Furthermore, a max pooling layer, a convolution layer with 128 filters, and another max pooling layer are followed, resulting in feature maps with the same

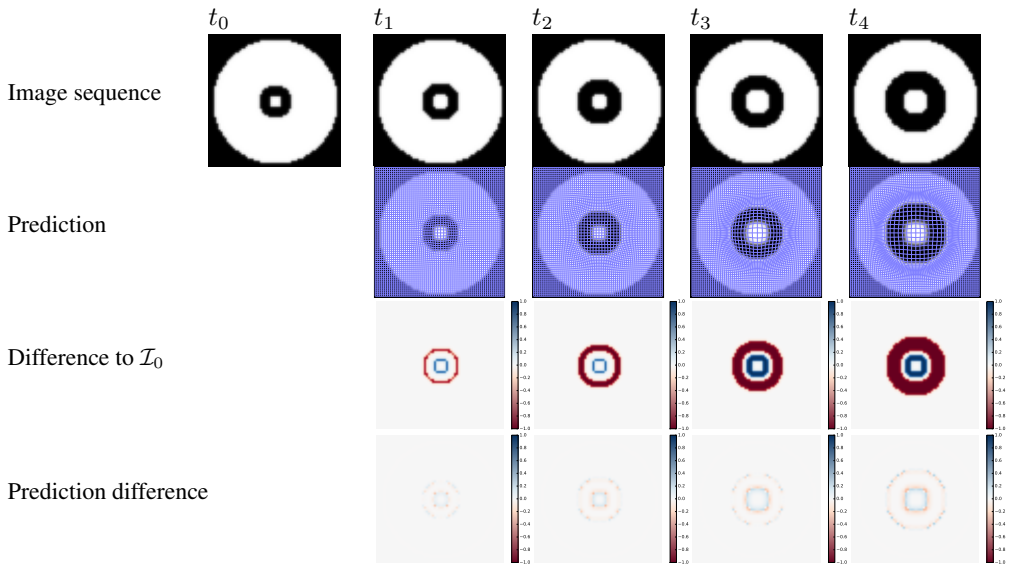

Figure 3: Prediction results for one sample image from the synthetic test set. The first row has the ground truths for the image sequence, following is our predicted sequence with deformation maps (the blue grids). The third row shows the image difference between the first image $\mathcal{I}_0$ and its follow-up images. The last row demonstrates the image difference between our prediction results and the corresponding ground-truth images.

size of the CNN image encoder for concatenation. The features learnt from this network are then forwarded to the decoder network.

CNN Vector Momentum Decoder. The decoder network aims to reconstruct a vector momentum sequence as the prediction $\{\hat{m}_{i,j}\}$ for the baseline image $I_{i,0}$ of a subject i. To achieve this, we use the inverse of the CNN image encoder architecture. It takes the features maps learned from the encoder network as input and predicts the next vector momentum in the sequence. This predicted momentum is then appended to the momentum sequence that is fed to the LSTM units to predict momentum at the further time steps. This recursive call is made until the last momentum in the sequence is predicted. This network consists of a up-sampling layer (upsampled by 2), a pair of convolution layers with 64 filters, which is followed by another up-sampling layer and a pair of convolution layers with 32 filters. The vector momentum output is generated by a convolution layer with 3 filters because of the x, y, and z components of the vector momentum.

2.3 Network Training and Prediction

Algorithm 1 depicts the workflow of our predictive network for image regression. It includes the preprocessing procedure discussed in Section 2.1, the training procedure for the network discussed in Section 2.2, and the prediction procedure for generating the image sequence. Although the LSTM network accepts a fixed number of images in the sequence, we can predict more images by taking the last prediction of the vector momentum and its corresponding image as the initial image and the initial momentum to continue the prediction procedure.

3 Experiments

Datasets and Experimental Settings. We have performed and evaluated experiments on both synthetic and real datasets. The synthetic data is a set of binary images of concentric circular rings like the bull eyes shown in Fig. 3. The radii of the concentric rings change by a constant factor but with some Gaussian noise. It has 52 image sequences with image size of 64×64 and each sequence has 5 time points. Among them, 40 momentum sequences are randomly selected for training, reserving 8 sequences for validation and 4 sequences for testing (total 16 images for prediction). The

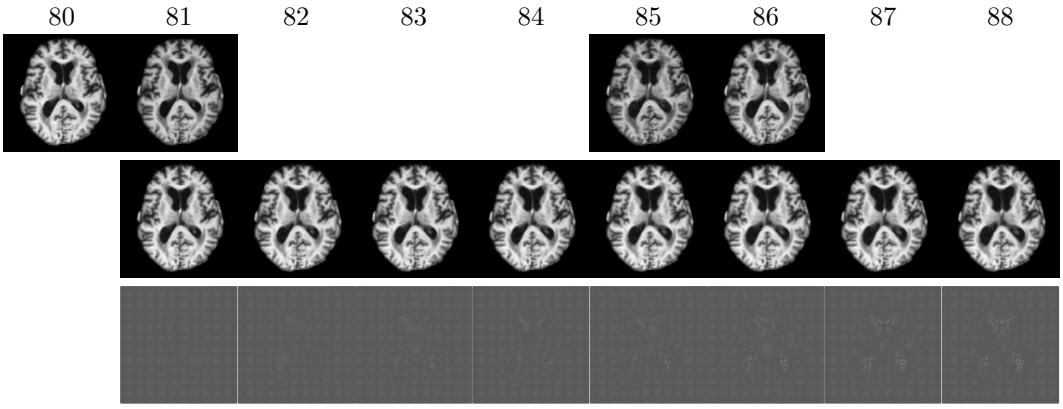

Figure 4: Prediction results of one sample subject from the test non-demented group. The first row has image scans of the subject at 80, 81, 85, and 86 years and all others are missing. The second row is our predicted image scans for this subject with the first scan at 80 years as input. The last row shows the corresponding deformation map for each predicted image.

computed vector momenta using LDDMM is of the form $64 \times 64 \times 3$ and one channel of the third dimension is zero. The image intensity is normalized within 0 and 1.

The real dataset includes 2D image slices of brain MRIs from the OASIS database. We have 136 subjects aged from 60 to 98 and each individuals was scanned at 2-5 time points with the same resampled resolution 128×128 and the voxel size of 1.25×1.25 mm^3. All images were preprocessed by down-sampling, skull-stripping, intensity normalization to the range [0,1], and co-registration with affine transformations. The vector momenta generated between pairs of these images are 128x128x3. We evaluate our model on the non-demented and demented groups separately. In particular, the non-demented group has 72 image sequences, 58 used for training, 7 for validation, and 7 for testing (total 12 images for prediction); and the demented group has 64 image sequences, 52 of them are used for training, 6 for validation, and 6 for testing (total 9 images for prediction).

The parameters for the L operator in LDDMM are set to [α, β, γ] = [0.01, 0.01, 0.001] and σ to 0.2.

Experimental Results. The prediction results of our model on the synthetic data are visually demonstrated in Fig. 3. We compute the image difference between each predicted image and its corresponding image in the sequence, as shown in the last row of Fig. 3, and compare it with the image changes relative to the first image in the sequence, as shown in the third row of Fig. 3. The dramatically reduced image difference indicates our prediction is promising. This is also validated by deformation maps overlapped on each image in the second row of Fig. 3, which shows the expanding changes have been correctly captured by our predictive model. The quantitative measure of the prediction results is included in Table 1. We compute the mean squared error (MSE) between an image and its prediction and estimate the mean and standard deviation for all predicted images in the test set of the synthetic data. The mean image difference is 8.1184e-4$\pm$5.3171e-4.

Figure 4 shows the prediction results of our model on one subject sample from the non-demented test group. This subject has scans

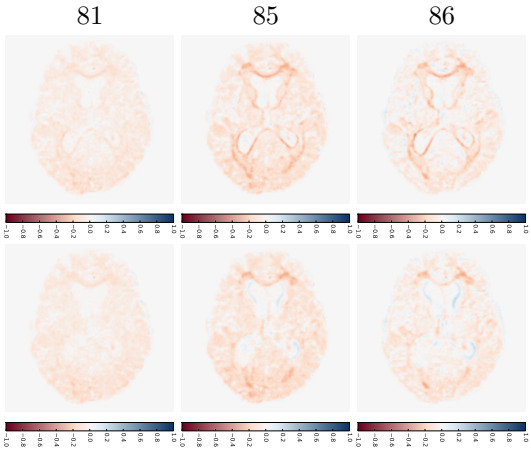

Figure 5: Difference plots for image scans and predictions in Fig. 4. The first row shows the image difference of image scans at 81, 85, and 86 years compared to the first scan of the subject at 80 years. The second row shows the difference between our predictions and images scanned at the corresponding years.

Synthetic data (1e-4)	Non-demented Group (1e-4)	Demented Group (1e-4)
8.1184 ± 5.3171	5.8616 ± 1.0703	6.1105 ± 1.9793

Table 1: The mean and standard deviation of mean squared errors over all images in each test group.

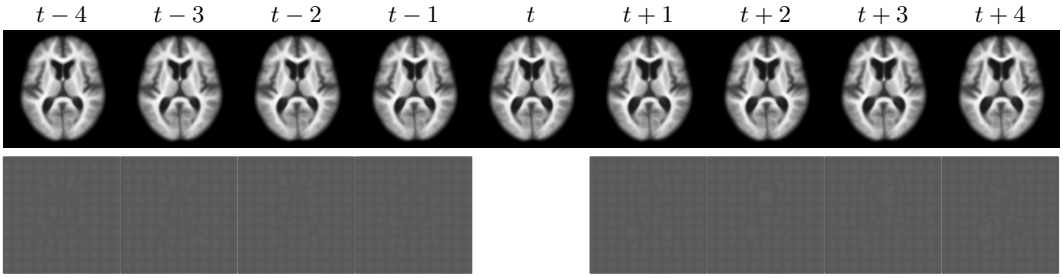

Figure 6: Prediction of forward and backward image sequences (top row) for the atlas of the demented group. The second row shows the corresponding deformation maps.

at 80, 81, 85, and 86 years old; therefore, there are several missing time points. We predict all the missing image scans and also predict future ones, as shown in the second row of Fig. 4. Since the brain changes are quite subtle, as shown in the first row of Fig. 5, we plot the deformation map for each time, as shown in the last row of Fig. 4. As we can see, the degeneration process of the ventricle in the brain has been captured, which is presented by expanding grids in the maps. Figure 5 shows the image difference between predicted images and their corresponding images on the second row and the difference is relatively smaller than the changes in the image sequence relative to the first scan, especially around the ventricle region. The means and standard deviations of the prediction difference for all images in the non-demented and demented groups are reported in Table 1, which are 5.8616e-4$\pm$1.0703e-4 and 6.1105e-4$\pm$1.9793e-4, respectively.

Our model can also estimate the group trajectory by predicting forward and backward image sequences for the atlas built for that group, as shown in Fig. 6. We predict vector momenta forward to generate future image scans and then use negative vector momenta at the atlas to predict backward and generate previous image scans. The image trajectory estimated for the demented group shows the time-varying changes of the ventricle in the brain.

4 Discussion and Conclusions

In this paper, we proposed a novel approach to predict time-varying medical image scans by integrating topologically-preserving image registration model (LDDMM) with deep neural networks. This model not only inherits the good properties from LDDMM and guarantees a sharp image prediction but also leverages the deep learning merits of learning from data, with need of parallel transport for specializing the prediction for a specific subject. Currently, we have the 2D implementation of our predictive model, but it can be straightforwardly extended to 3D images.

The limitation of our method lies in the LSTM network, which mainly focuses on capturing linear changes in the image sequence, because of its shared weights among units of a layer for all time points. In the future, we consider of another recurrent neural network or improve LSTM to capture non-linear changes. In addition, the network cannot deal with missing correspondences among images, like a tumor appearing or disappearing in the brain images. This is caused by the LDDMM framework we used, which cannot handle image registrations with missing correspondences. To address this issue, we can modify the current LDDMM framework with the metamorphic image registration [19], which was developed under the LDDMM framework.

References

[1] Marc Niethammer, Yang Huang, and François-Xavier Vialard. Geodesic regression for image time-series. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 655–662. Springer, 2011.

[2] M Faisal Beg, Michael I Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International journal of computer vision*, 61(2):139–157, 2005.

[3] Jacob Hinkle, P Thomas Fletcher, and Sarang Joshi. Intrinsic polynomials for regression on riemannian manifolds. *Journal of Mathematical Imaging and Vision*, 50(1-2):32–52, 2014.

[4] Nikhil Singh, François-Xavier Vialard, and Marc Niethammer. Splines for diffeomorphisms. *Medical image analysis*, 25(1):56–71, 2015.

[5] Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P Thomas Fletcher. Hierarchical geodesic models in diffeomorphisms. *International Journal of Computer Vision*, 117(1):70–92, 2016.

[6] Yi Hong, Polina Golland, and Miaomiao Zhang. Fast geodesic regression for population-based image analysis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2017.

[7] Yi Hong, Yundi Shi, Martin Styner, Mar Sanchez, and Marc Niethammer. Simple geodesic regression for image time-series. In *International Workshop on Biomedical Image Registration*, pages 11–20. Springer, 2012.

[8] Kristen M Campbell and P Thomas Fletcher. Efficient parallel transport in the group of diffeomorphisms via reduction to the lie algebra. In *Graphs in Biomedical Image Analysis, Computational Anatomy and Imaging Genetics*, pages 186–198. Springer, 2017.

[9] Brad C Davis, P Thomas Fletcher, Elizabeth Bullitt, and Sarang Joshi. Population shape regression from random design data. *International journal of computer vision*, 90(2):255–266, 2010.

[10] Monami Banerjee, Rudrasis Chakraborty, Edward Ofori, Michael S Okun, David E Viallancourt, and Baba C Vemuri. A nonlinear regression technique for manifold valued data with applications to medical image analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4424–4432, 2016.

[11] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015.

[12] William Lotter, Gabriel Kreiman, and David Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104*, 2016.

[13] Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440*, 2015.

[14] Nikhil Singh, Jacob Hinkle, Sarang Joshi, and P Thomas Fletcher. A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction. In *Biomedical Imaging (ISBI), 2013 IEEE 10th International Symposium on*, pages 1219–1222. IEEE, 2013.

[15] Zhipeng Ding, Greg Fleishman, Xiao Yang, Paul Thompson, Roland Kwitt, Marc Niethammer, Alzheimer's Disease Neuroimaging Initiative, et al. Fast predictive simple geodesic regression. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 267–275. Springer, 2017.

[16] Xiao Yang, Roland Kwitt, Martin Styner, and Marc Niethammer. Quicksilver: Fast predictive image registration–a deep learning approach. *NeuroImage*, 158:378–396, 2017.

[17] François-Xavier Vialard, Laurent Risser, Daniel Rueckert, and Colin J Cotter. Diffeomorphic 3d image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97(2):229–241, 2012.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015.

[19] Yi Hong, Sarang Joshi, Mar Sanchez, Martin Styner, and Marc Niethammer. Metamorphic geodesic regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 197–205. Springer, 2012.