Learning Generalizable Recurrent Neural Networks from Small Task-fMRI Datasets

Nicha C. Dvornek[1(✉)], Daniel Yang[3], Pamela Ventola[2], and James S. Duncan[1,4,5]

[1] Department of Radiology & Biomedical Imaging, Yale School of Medicine, New Haven, CT, USA
nicha.dvornek@yale.edu
[2] Child Study Center, Yale School of Medicine, New Haven, CT, USA
[3] Autism and Neurodevelopmental Disorders Institute, George Washington University and Children's National Health System, Washington, DC, USA
[4] Department of Biomedical Engineering, Yale University, New Haven, CT, USA
[5] Department of Electrical Engineering, Yale University, New Haven, CT, USA

Abstract. Deep learning has become the new state-of-the-art for many problems in image analysis. However, large datasets are often required for such deep networks to learn effectively. This poses a difficult challenge for many medical image analysis problems in which only a small number of subjects are available, e.g., patients undergoing a new treatment. In this work, we propose a number of approaches for learning generalizable recurrent neural networks from smaller task-fMRI datasets: (1) a resampling method for ROI-based fMRI analysis to create augmented data; (2) inclusion of a small number of non-imaging variables to provide subject-specific initialization of the recurrent neural network; and (3) selection of the most generalizable model from multiple reinitialized training runs using criteria based on only training loss. Using cross-validation to assess model performance, we demonstrate the effectiveness of the proposed methods to train recurrent neural networks from small datasets to predict treatment outcome for children with autism spectrum disorder ($N = 21$) and classify autistic vs. typical control subjects ($N = 40$) from task-fMRI scans.

1 Introduction

Deep learning approaches are quickly becoming the machine learning technique of choice for many medical image analysis problems, e.g., image classification, segmentation, and registration [12]. The deep neural networks have a large capacity to learn directly from raw images. However, it is well known that these popular methods can quickly overfit the data, resulting in poor generalization. Thus,

This work was supported by NIH grants R01MH100028 and R01NS035193.

A. F. Frangi et al. (Eds.): MICCAI 2018, LNCS 11072, pp. 329–337, 2018.
https://doi.org/10.1007/978-3-030-00931-1_38

learning useful models generally requires training on very large datasets and using proper model validation techniques.

The large data requirement poses a challenge in analyzing many medical imaging datasets, in which only a small number of subjects may be available. For example, it may not be feasible to gather large amounts of data when studying a specific disease population or a new experimental treatment, or it may be difficult to obtain expert manual annotations of large datasets for training. Recent trends toward open science and data sharing have made larger medical imaging datasets more widely available, e.g., the ABIDE dataset for autism [2]; however, creating large datasets for every disease and medical imaging problem is clearly not possible. While some medical image analysis problems can handle smaller datasets by using patch-based approaches (e.g., in image segmentation [12]) to augment the amount of data, such methods are not as suitable for analyzing neurological data from functional magnetic resonance imaging (fMRI).

In this paper, we propose new strategies that facilitate learning more generalizable neural network models from small fMRI datasets. We first adopt a recurrent neural network with long short-term memory (LSTM) to generate predictions from a whole-brain parcellation of fMRI data. We then use resampling approaches to generate multiple summary time-series for each region in the parcellation, augmenting the original dataset. Next, we utilize available non-imaging variables to provide subject-specific initialization of the LSTM network. Finally, we describe a criteria for selecting the most generalizable model from many training instances on the same data using only training loss, allowing all available data to be used for model training. We apply the proposed strategies and compare them to other approaches to learn from task-fMRI for two small data examples: (1) a regression problem of predicting treatment outcome from 21 children with autism spectrum disorder (ASD), and (2) a classification problem of identifying autistic children vs. typical controls from a dataset of 40 subjects.

2 Methods

2.1 Base LSTM Architecture for fMRI

LSTMs and related architectures are designed to learn long-term dependencies in time-series data [7]. They have recently been applied to fMRI for modeling brain dynamics [6] and for classification [3]. In addition to the time dependent nature of the model, LSTMs are a nice neural network model specifically for small fMRI datasets, since an "unrolled" LSTM with T timesteps can be thought of as a deep network with T layers that share the same parameters across all the layers. This likely considerably reduces the number of model parameters that need to be learned compared to other standard deep neural network architectures.

Standard fMRI whole-brain analysis involves first summarizing the data in a number of regions of interest (ROIs) according to some brain parcellation. While deep networks are able to learn from raw image inputs, the ROI approach is beneficial in our case of dealing with smaller fMRI datasets, as fMRI data is very noisy and the ROI representation greatly reduces the input data dimension.

Our base LSTM architecture is based on the model in [3], with added regularization and slight changes for regression vs. binary classification. The summary time-series from the ROIs are used as input to an LSTM. For regression, we pass the output from the LSTM at the last timestep to a dense layer to produce the predicted value (Fig. 1(a), blue path); thus, the entire task-fMRI sequence is analyzed before providing a final prediction. For classification, we more closely follow the network in [3]; the LSTM output from every timestep is passed to a shared dense layer with a single node, followed by mean pooling across time and a sigmoid activation function to produce the classification probability (Fig. 1(b), blue path). During training, we include dropout of the LSTM weights [5] and add dropout (with probability 0.5) between the LSTM output and dense layer.

2.2 Data Augmentation by Resampling

Standard data augmentation techniques for neural networks to learn from image data include using random croppings and random rotations of the images. However, our LSTM network is designed to use the time-series from the brain ROIs as inputs, and such augmentation techniques are not appropriate for fMRI. We could perform random cropping along the time dimension, but LSTMs learn best from long sequences. Another generic approach is to inject random noise into the inputs [16]; however, it is unclear how to choose the best noise model and associated parameters, and while such approaches may slow down overfitting, it may not be representative of the variation in the fMRI data.

We instead propose a resampling approach to augment the data. Traditional ROI analysis extracts the mean time-series calculated from all voxels in the ROI. To inject variation to the ROI time-series, we propose sampling only a subset of the ROI voxels or sampling all voxels with replacement (bootstrapping) and using the average of the sampled data to summarize the time-series for the ROI.

2.3 LSTM Initialization with Non-imaging Variables

An LSTM cell contains two state vectors, the hidden state (i.e., output) h_t and the cell state c_t. The state of an LSTM at time t depends on the current input x_t and the cell state from the previous timestep h_{t-1} and c_{t-1}:

$$g_t = \sigma\left(W_g x_t + U_g h_{t-1} + b_g\right), \text{with } g \in \{i, f, o\} \tag{1}$$

$$\tilde{c_t} = \tanh\left(W_c x_t + U_c h_{t-1} + b_c\right) \tag{2}$$

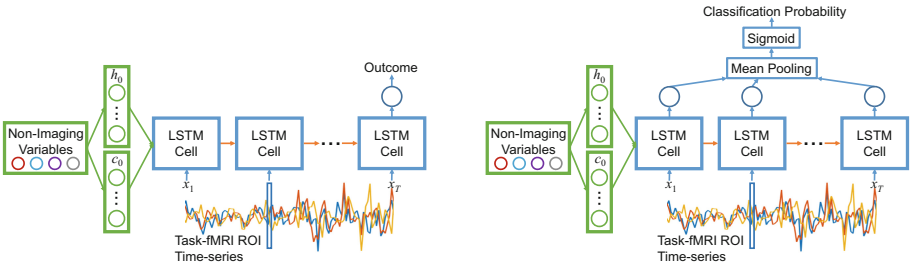
$$c_t = i_t * \tilde{c_t} + f_t * c_{t-1} \tag{3}$$

$$h_t = o_t * \tanh\left(c_t\right) \tag{4}$$

where i, f, and o represent input, forget, and output gates, $\tilde{c_t}$ is the current estimated cell state, and W, U, and b are the LSTM model parameters.

Unless otherwise specified, the initial state of the LSTM is set to zeros, $h_0 = c_0 = \mathbf{0}$. Simple non-imaging subject information (e.g., age) is often available. We propose initializing the LSTM by feeding such non-imaging information into 2

dense layers, whose outputs are the initial hidden and cell states (Fig. 1, green path). Such initialization approaches have been proposed in other domains, e.g., an LSTM model to generate an image caption was initialized on image features extracted via a convolutional neural network [10]. In our small data setting, conditioning the LSTM on subject-specific parameters helps to incorporate non-imaging variation across subjects with just a small increase in model parameters, unlike other multi-modal fusion techniques for neural networks [4,13].



(a) Regression Network (b) Classification Network

Fig. 1. LSTM networks with initialization using non-imaging variables.

2.4 Model Selection from Training Loss

Neural network training (in the non transfer-learning case) is generally performed using random initialization of model weights. With large amounts of data, several training runs can be performed with different initializations, and a validation dataset can be set aside to assist with choosing the best trained model. However, with small datasets, we would prefer to use all available data for training. Furthermore, splitting off a small validation set is likely not representative of the test data and thus is not appropriate for model selection.

We propose choosing the best model from several initializations based on the recorded training loss curve. Rather than choosing the model with the lowest loss, which is likely to be the most overfit to the small dataset, we choose the model that fits slowest to the data. We quantify this criteria with the following:

$$\hat{M} = \arg\max_{M} \left[median\left(\triangle L_{M,s}\right) \frac{1}{L_M\left(0\right) \times s} \right], \tag{5}$$

where L_M is the training loss curve for model M, $L_M\left(0\right)$ is the loss after epoch 0, $\triangle L_{M,s}$ are the first differences of the loss curve from epoch 0 to stopping epoch s, and s is the first epoch such that $L_M\left(s\right) < {}^{L_M(0)}/_{e}$. Thus, the criteria is looking for the model that learns slowest, measured by the median of the first differences over the epochs up to epoch s, weighted by the initial loss and the number of epochs to reduce the loss to ${}^{1}/_{e}$ of the initial loss (borrowing the idea of relaxation time). We only look at the first differences up to ${}^{1}/_{e}$ since we are more interested

in how fast the model fits the data in earlier epochs. Furthermore, the training curve will likely have a very long flat tail due to overfitting to the training set, making it difficult to measure differences in convergence. We scale our criteria by the initial loss, since given two curves with the same rate of convergence but with different initial losses, we would rather choose the model with the higher initial loss, signifying a worse initial fit and overall slower learning of the model. Finally, we scale by the number of epochs for the signal to decay ("relaxation time"), since given two curves with similar convergence measured by the other two metrics, we want the model that takes longer to minimize the loss.

3 Experiments

3.1 Data and Preprocessing

Data was acquired from 21 children with ASD (ages 6.05 ± 1.24 years) and 19 typically-developing controls (TC) (ages 6.42 ± 1.29 years). Each subject underwent a T1-weighted MP-RAGE structural MRI ($1 \times 1 \times 1$ mm^3 voxel size) and BOLD T2*-weighted fMRI sequence ($3.44 \times 3.44 \times 4.00$ mm^3 voxel size) acquired during a biological motion perception task [9]. The fMRI paradigm involved viewing point light animations of coherent and scrambled biological motion in a block design ($\sim$24 s per block, $\sim$5 min scan). Non-imaging information collected included age, sex, IQ, and score on the Social Responsiveness Scale (SRS), 2nd edition [1]. SRS measures severity of social impairment in autism; lower scores signify better social ability. ASD subjects then underwent 16 weeks of Pivotal Response Treatment [11], a behavioral therapy for ASD. SRS score was measured again at the end of treatment.

Images were preprocessed in FSL [8] using the pipeline by Pruim et al. [14], which included motion correction, interleaved slice timing correction, brain extraction, 4D mean intensity normalization, spatial smoothing (5 mm FWHM), data denoising via ICA-AROMA [14], nuisance regression using white matter and cerebrospinal fluid, and high-pass temporal filtering (100 s). Functional MRI were aligned to the standard MNI brain with the aide of the structural MRI. The AAL atlas [15] was applied, resulting in 90 cerebral ROIs from which summary time-series (156 timepoints) were extracted and used as input to the LSTM. Since fMRI absolute signal varies greatly across the brain, each summary time-series was standardized (subtracted mean, divided by standard deviation). The data for each non-imaging variable were normalized to range $[-1, 1]$.

3.2 Regression Example: Prediction of Treatment Outcome

We investigated the effectiveness of the proposed learning strategies on the following regression problem: to predict the treatment outcome (i.e., percent change in SRS) for the 21 children with ASD from baseline information. Leave-one-out cross-validation (train on 20, test 1) was used to assess model performance. Mean squared error (MSE), standard deviation (SD) of the squared error, and Pearson's correlation coefficient (r) between predicted and true treatment outcome

were computed from cross-validation folds. Paired one-tailed t-tests were used to compare the squared errors, and p-values for r provided evidence for non-zero correlation, with a significance level of 0.05. Neural networks were implemented and trained in Keras using the MSE loss function, adadelta optimizer, 8 hidden LSTM units, a maximum of 100 epochs with early stopping (patience of 5 epochs monitoring training loss), and a batch size of 32 unless otherwise specified.

We first directly trained the LSTM network on the 21 fMRI datasets. We varied the batch size (2, 5, 10, 20) to try to improve learning. The best result is shown in Table 1a ("Original"); however, errors between the best result and other batch sizes were not significantly different, and correlations were insignificant.

Table 1. Results for predicting treatment outcome.

(a) Data augmentation approaches.

Dataset	MSE (SD)	r	p_r
Original	0.097 (0.160)	0.35	0.1204
Repeat	0.035 (0.049)	0.45	0.0415
Low Noise	0.034 (0.037)*	0.47	0.0324
High Noise	0.029 (0.029)*	0.59	0.0050
Sample 10	0.029 (0.034)*	0.58	0.0058
Sample 50	0.034 (0.036)*	0.47	0.0313
Sample 250	0.030 (0.026)*	0.55	0.0101
Bootstrap	0.031 (0.041)*	0.53	0.0129

*Significantly better than original dataset.

(b) Non-imaging data and model selection.

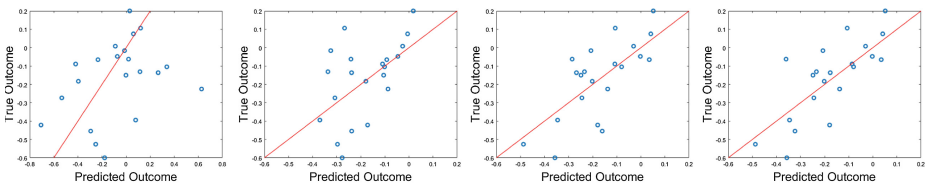
Dataset	MSE (SD)	r	p_r
Bootstrap (BS)	0.031 (0.041)	0.53	0.0129
BS + Non-Imaging	0.020 (0.025)	0.73	0.0002
BS + Top Fusion	0.035 (0.037)	0.46	0.0339
BS + Model Bag	0.032 (0.037)	0.51	0.0175
BS + Model Select	0.028 (0.032) †	0.60	0.0044
BS + Non-Imaging + Model Bag	0.024 (0.029) †	0.66	0.0011
BS + Non-Imaging + Model Select	0.018 (0.025)^†	0.77	<0.0001

^Significantly better than bootstrap dataset.
†Significantly better than one individual model.

We then compared the following data augmentation techniques: (1) repeating the data ("Repeat"), (2) standard noise injection by adding zero-mean Gaussian noise with SD equal to SD of the time-series divided by 10 ("Low Noise") or 2 ("High Noise"), and (3) the proposed resampling approach of randomly sampling 10, 50, or 250 voxels without replacement or bootstrap sampling all voxels from each ROI to compute the summary time-series. We repeated the augmentation approaches 50 times per subject, resulting in 1050 samples. Results are shown in Table 1a. Simply repeating the data resulted in significant correlation, although MSE did not significantly improve. All other augmentation approaches produced significant correlation as well as significantly reduced the MSE. While the high noise augmentation nominally resulted in the highest correlation, there were no significant differences between any noise and sampling methods.

Since errors were not significantly different and the bootstrap sampling does not require any parameter selection, we tested the remaining learning strategies on only the bootstrap-augmented dataset (Table 1b). Initializing the LSTM with non-imaging data dramatically improved the correlation and reduced the MSE by 35% (just missing significance with $p = 0.0572$), at the cost of only a 1% increase in number of parameters. Applying a standard multimodal fusion approach to combine the final fMRI score and non-imaging data in a dense layer, also increasing the number of parameters by 1%, results in worse performance ("Top Fusion"), demonstrating the benefit of our LSTM initialization method.

We tested our model selection approach by assessing the training curves from 2 separate runs, and compared this to averaging the predictions from the 2 runs (bagging). We applied these approaches to the bootstrap dataset and the bootstrap with non-imaging model. Model bagging did not produce significantly lower errors compared to the individual models for the bootstrap dataset. Our model selection approach resulted in significantly lower MSE compared to at least one of the individual models. Furthermore, applying all three of our proposed learning strategies resulted in significantly more accurate predictions compared to data augmentation alone, with the highest correlation with the true outcomes. The effect of adding each proposed learning strategy is illustrated in Fig. 2.



(a) (b) (c) (d)

Fig. 2. Plots of true vs. predicted treatment outcome after applying each proposed learning strategy. Perfect predictions would fall on red reference line. (a) Original data. (b) Data augmentation with bootstrap resampling. (c) Bootstrap resampling and LSTM initialization with non-imaging data. (d) Bootstrap resampling, inclusion of non-imaging data, and model selection based on training loss criteria in (5).

3.3 Classification Example: Autism vs. Typical Control

We tested the proposed learning strategies on training the LSTM network to classify the 21 ASD and 19 TC subjects (52.5% ASD subjects). Ten-fold cross-validation (train on 36, test 4) was repeated 10 times, and performance of different methods were measured using mean and standard deviation of the cross-validation accuracy, true positive rate (TPR), and true negative rate (TNR). Paired one-tailed t-tests were used to compare cross-validation performance between different methods, with a significance level of 0.05. Training was run with similar Keras setup as above, with maximum number of epochs reduced to 20.

Results quantifying the effects of the proposed learning strategies are shown in Table 2. Learning from the original, non-augmented sample results in chance accuracy. Applying bootstrap resampling (50 resamples, resulting in 2000 total samples) significantly improves the accuracy and TPR. Using non-imaging variables to set the initial LSTM state further improved all performance measures, with significant differences compared to using the original dataset. Finally, additionally including model selection produced the best performing model.

Table 2. Results for classifying ASD vs. TC subjects.

Dataset	Mean (SD) Accuracy (%)	Mean (SD) TPR (%)	Mean (SD) TNR (%)
Original	51.8 (3.3)	56.1 (13.3)	55.1 (12.4)
Bootstrap	64.5 (5.1)*	70.7 (7.3)*	60.9 (11.1)
Bootstrap + Non-imaging	67.5 (6.7)*	72.2 (9.2)*	64.6 (6.3)*
Bootstrap + Non-imaging + Model select	69.8 (5.5)*^†	75.1 (8.4)*	65.5 (6.8)*

*Significantly better than original dataset. ^Significantly better than bootstrap dataset. †Significantly better than at least one individual model.

4 Conclusions

In this work, we presented strategies for training LSTMs on small datasets and demonstrated their effectiveness in learning better generalized models. Our methods for facilitating learning included a data augmentation approach specific to ROI-based analysis, incorporation of subject-specific variations by initializing the LSTM based on each subject's non-imaging parameters, and model selection based on training loss criteria alone to maximize the amount of data available for training. Regression and classification learning from 2 small task-fMRI datasets showed that while naïve training of the LSTM was unable to learn useful models, combining the proposed learning strategies resulted in the successful training of more generalizable LSTMs.

References

1. Constantino, J.N., Gruber, C.P.: The Social Responsiveness Scale (SRS-2), 2nd edn. Western Psychological Services, Torrance (2012)
2. Di Martino, A., et al.: The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol. Psychiatry **19**, 659–667 (2014)
3. Dvornek, N.C., Ventola, P., Pelphrey, K.A., Duncan, J.S.: Identifying autism from resting-state fMRI using long short-term memory networks. In: Wang, Q., Shi, Y., Suk, H.-I., Suzuki, K. (eds.) MLMI 2017. LNCS, vol. 10541, pp. 362–370. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67389-9_42
4. Dvornek, N.C., Ventola, P., Duncan, J.S.: Combining phenotypic and resting-state fMRI data for autism classification with recurrent neural networks. In: ISBI (2018)
5. Gal, Y., Ghahramani, Z.: A theoretically grounded application of dropout in recurrent neural networks. In: NIPS (2016)
6. Güçlü, U., van Gerven, M.A.J.: Modeling the dynamics of human brain activity with recurrent neural networks. Front. Comput. Neurosci. **11**, 7 (2017)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
8. Jenkinson, M., Beckmann, C.F., Behrens, T.E., Woolrich, M.W., Smith, S.M.: FSL. NeuroImage **62**, 782–790 (2012)

9. Kaiser, M., et al.: Neural signatures of autism. Proc. Natl. Acad. Sci. U.S.A. **107**(49), 21223–21228 (2010)
10. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 664–676 (2017)
11. Koegel, R., Koegel, L.: Pivotal Response Treatments for Autism: Communication, Social, and Academic Development. Brookes Publishing Company, Baltimore (2006)
12. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–68 (2017)
13. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: The 28th International Conference on Machine Learning (2011)
14. Pruim, R.H., Mennes, M., van Rooij, D., Ller, A., Buitelaar, J.K., Beckmann, C.F.: ICA-AROMA: a robust ICA-based strategy for removing motion artifacts from fMRI data. NeuroImage **112**, 267–277 (2015)
15. Tzourio-Mazoyera, N., et al.: Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage **15**, 273–289 (2002)
16. Zur, R.M., Jiang, Y., Pesce, L.L., Drukker, K.: Noise injection for training artificial neural networks: a comparison with weight decay and early stopping. Med. Phys. **36**, 4810–4818 (2009)