A scaling study of self-supervised auto-regressive modelling of fMRI time series and performance on downstream sex prediction in the UK biobank sample

Hao-Ting Wang (wang.hao-ting@criugm.qc.ca)
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, QC Canada

François Paugam (francois.paugam@umontreal.ca,)
Department of Psychology, Universitaire de Montréal, Montréal, QC Canada

Nicolas Farrugia (nicolas.farrugia@imt-atlantique.fr)
IMT Atlantique, Brest, France

Pierre Bellec (pierre.bellec@criugm.qc.ca)
Centre de recherche de l'Institut universitaire de gériatrie de Montréal, Montréal, QC Canada
Department of Psychology, Universitaire de Montréal Montréal, QC Canada
Mila, Universitaire de Montréal, Montréal, QC, Canada

Abstract

Limited data availability restricts how deep learning techniques can model patterns of brain activity. This study investigates the impact of training set size on the capacity of autoregressive graph convolutional networks (GCNs) to learn predictive features from fMRI data. GCN's graph structure, few parameters and interpretable features make it a good candidate to model complex brain dynamics. Using a large fMRI dataset, we assessed how dataset size impacts (1) autoregressive GCN's ability to predict future fMRI time points from past time points, and (2) the predictive value of GCN's learned features on a downstream sex prediction task. Our findings show performance saturation at a sample size of 8000 subjects for both the autoregressive and the downstream task, highlighting the model's ability to capture relevant brain signals. Standard deviation pooling from GCN layer weights emerged as the most predictive feature on the downstream task. These results motivate further exploration into more complex model architectures to achieve gains in performance.

Keywords: fMRI; auto-regressive model; functional connectivity; foundation model

Introduction

While deep learning offers powerful tools to model the complexity of brain activity, its integration to brain imaging data analysis pipelines is hindered by the small sample size of most studies. This data scarcity restricts deep learning models that are typically trained on enormous datasets. The current study trains autoregressive graph convolutional networks (GCNs; Wu et al.) on functional magnetic resonance imaging (fMRI) data, and investigates the impact of sample size on the reliability of model predictions. In the context of fMRI analyses, GCNs are well suited to model complex brain dynamics due to their graph structure, few parameters and interpretable features. GCNs can be combined with an autoregressive model to learn to predict future brain activity patterns (fMRI BOLD frames) from past time points. These features learned from the brain can then be applied to downstream tasks, including categorical predictions.

This research has two objectives:

1. Apply scaling to the sample size to determine the amount of data needed to achieve robust brain signal reconstruction with the autoregressive GCN model.
2. Investigate how features learned during GCN training perform on downstream prediction tasks as a function of sample size.

These findings will inform future studies by providing insights into the optimal sample size necessary to leverage deep learning techniques in fMRI research.

Methods

Model

Autoregressive models predict future signal given the values of previous time points. The task can be formalised as the following equation:

$$X(t+l) = f(\{X(i)|t-k < i <= t\}) + \varepsilon(t+l)$$

where X is the BOLD time series, t the time index, l the lag between the predicted time point and the last time point in the input, f the model, k the number of past time points used as input and ε the error term to minimize. In this study, we focused on models trained for prediction at lag $l = 1$.

The choice of model architecture was based on an autoregressive model benchmark based on fMRI data (Paugam, Pinsard, Lajoie, & Bellec, 2023). We use the Chebnet (Defferrard, Bresson, & Vandergheynst, 2016), a type of graph convolutional neural network (Wu et al., 2021), whose nodes correspond to brain parcels and node signals correspond to the parcels' time series For each subject, the binary adjacency matrix used to compute the graph convolutions was estimated by selecting the 10% most correlated pairs of parcels (Pearson's correlation) on the average functional connectome derived from the training data.

Hyper parameter tuning

We optimized the GCN's autoregressive performance (R^2) on 20793 subjects (80% of the total 25992 subjects) with a 75%-25% training-validation split, using a random search over the following hyper-parameters: number of ChebNet layers (3, 6, 13; fixed 8 x 6 size), pooling MLP layer (16-8-1 structure or single node), learning rate (1e-4 to 0.3 range with 1e-6 to 1 threshold), epochs (16-24), dropout (0-0.3), batch size (128-256) and the number of time points (12-32) considered for prediction.

Dataset

We preprocessed fMRI data from 38998 subjects from the UK BioBank (Sudlow et al., 2015) resting-stated task with fMRIPrepLTS 20.2.7 (Esteban et al.; RRID:SCR_016216), and excluded subjects with excessive motion, resulting in 25992 subjects for the final analysis. The preprocessed data was denoised with load_confounds (Wang et al.), and fMRI time series were extracted from 197 brain regions defined by the MIST parcellation (Urchs et al., 2019). The fMRI data was acquired at TR=0.735s, but time series were decimated by keeping only every 4^{th} sample to estimate potential model behaviour on more typical fMRI datasets (TR≅2.5s).

Scaling experiment

The configuration identified with hyperparameter search was used for the scaling experiment.

To investigate the impact of sample size on performance, the experiment was performed on datasets of varying sizes, range from 100 subjects to the full dataset containing 25,992 subjects: 100, 250, 500, 1000, 2000, 3000, 4000, 5000, 6000, 8000, 10000, 16000, and 20000. For each sample size, the training process was replicated using 6 random seeds[1] to account for variations in model performance. The data was split into 60% training the autoregressive model, 20% for validation of the autoregressive model, 20% for downstream prediction.

Downstream tasks Trained GCN features were extracted for a downstream sex prediction task. Features were fed into a Support Vector Machine (SVM) classifier with l2 penalty to predict the sex label (male or female) of each subject. The data were split into an 80% training set and a 20% testing set for model evaluation to assess the model's capacity to generalize to unseen data.

The following features, which were meant to capture information about the underlying brain connectivity, were selected as predictors.

- The model's predictions (R^2 map) at the next time point (t+1) across all brain regions, which reflects how well it explains variance in future brain activity.

- The model's average R^2, which summarizes the model's overall explanatory power for future brain activity.

- Average pooling, standard deviation pooling and max pooling, three summary statistics extracted from the convolutional layers weights that capture activation patterns in the model's learned filters, potentially reflecting underlying brain connectivity patterns.

- A 1D convolution operation on the weights of the convolutional layers.

For comparison, the functional connectome, which is well known to scale by number of subjects, was also included as a baseline feature.

Results

The best performing model used a ChebNet architecture with 6 convolutional layers of size 8 x 6 and MLP layers with 16-8-1 nodes trained for 18 epochs (dropout = 0.022, batch size = 225, lr = 0.05, threshold = 0.411) using a sequence length of 29. The model's performance, as measured by validation set R^2, exhibited a saturation effect with increasing sample size, achieving a plateau at approximately 8000 subjects, corresponding to 4800 subjects in the training set and 1600 in the validation set (see Figure 1). The mean validation R^2 is 0.186, which indicates a poor fit. This does not affect the plan to investigate the downstream predictions as the work aims to establish a benchmark for exploring more complex architecture.

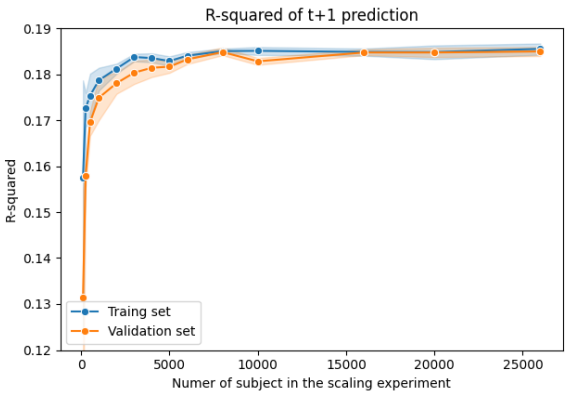

Figure 1: Autoregresive model performance scaled by number of subjects in each experiment.

The downstream sex prediction task revealed a similar pattern to the scaling experiment (see Figure 2), demonstrating that the model's ability to extract meaningful information from fMRI data stabilizes around a sample size of 8000 subjects, corresponding to 6400 subjects for autoregressive model training and evaluation, tested on 1600 subjects as the hold out set for downstream task. The prediction accuracy from the connectome was 92.3% on average and the best feature from the GCN was the standard deviation pooling, with an accuracy of 79.3%.

[1] 0, 1, 2, 4, 8, and 42

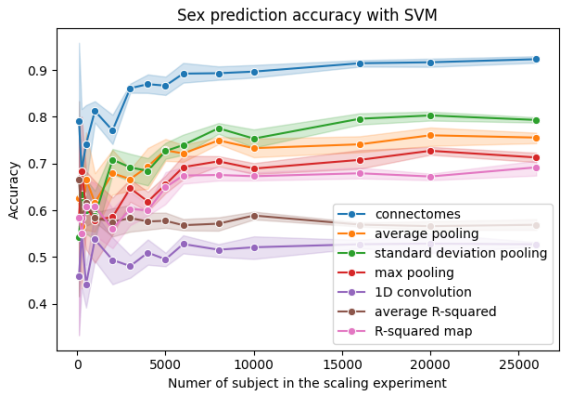

Figure 2: Sex prediction accuracy scaled by number of subjects in each experiment.

Discussion

A relatively simple autoregressive GCN model achieved performance saturation for fMRI analysis at a sample size of 8000 subjects. This saturation effect held true for both the core model predictions and downstream sex prediction tasks, with standard deviation pooling being the most effective feature extraction method. Despite the poor model performance, the extracted features achieved around 80% accuracy rate on sex prediction. This is still below the baseline established by functional connectomes. These findings motivate exploration of more complex architectures for potentially even better performance.

Acknowledgements

The project was supported by the following fundings: Digital Alliance Canada Resource Allocation Competition (RAC 1827 and RAC 4455) to PB, Institut de Valorisation des Données projets de recherche stratégiques (IVADO PFR3) to PB, and Canadian Consortium on Neurodegeneration in Aging (CCNA; team 9 ”discovering new biomarkers”) to PB, the Courtois Foundation to PB, and Institut national de recherche en sciences et technologies du numérique (INRIA; Programme Équipes Associées - NeuroMind Team DRI-012229) to PB. HTW is supported by IVADO postdoc fellowship. FP is supported by Courtois Foundation Neuromod Project. PB was funded by Fonds de Recherche du Québec - Santé.

References

Defferrard, M., Bresson, X., & Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *CoRR*, *abs/1606.09375*. doi: 10.48550/arXiv.1606.09375

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... Gorgolewski, K. J. (2019, January). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), 111–116. doi: 10.1038/s41592-018-0235-4

Paugam, F., Pinsard, B., Lajoie, G., & Bellec, P. (2023). *A benchmark of individual auto-regressive models in a massive fmri dataset.* PsyArXiv. doi: 10.31234/osf.io/pvx3d

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., ... Collins, R. (2015, 03). Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, *12*(3), 1-10. doi: 10.1371/journal.pmed.1001779

Urchs, S., Armoza, J., Moreau, C., Benhajali, Y., St-Aubin, J., Orban, P., & Bellec, P. (2019, March). MIST: A multiresolution parcellation of functional brain networks. *MNI Open Research*, *1*, 3. doi: 10.12688/mniopenres.12767.2

Wang, H.-T., Meisler, S. L., Sharmarke, H., Clarke, N., Gensollen, N., Markiewicz, C. J., ... Bellec, P. (2024, 03). Continuous evaluation of denoising strategies in resting-state fmri connectivity using fmriprep and nilearn. *PLOS Computational Biology*, *20*(3), 1-32. doi: 10.1371/journal.pcbi.1011942

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4-24. doi: 10.1109/TNNLS.2020.2978386