Anatomically-Informed Data Augmentation for Functional MRI with Applications to Deep Learning

Kevin P. Nguyen, Cherise Chin Fatt, Alex Treacher, Cooper Mellema, Madhukar H. Trivedi, and Albert Montillo

University of Texas Southwestern Medical Center, Dallas, TX, USA

Keywords: data augmentation, neuroimaging, fMRI, depression, deep learning

ABSTRACT

The application of deep learning to build accurate predictive models from functional neuroimaging data is often hindered by limited dataset sizes. Though data augmentation can help mitigate such training obstacles, most data augmentation methods have been developed for natural images as in computer vision tasks such as CIFAR, not for medical images. This work helps to fills in this gap by proposing a method for generating new functional Magnetic Resonance Images (fMRI) with realistic brain morphology. This method is tested on a challenging task of predicting antidepressant treatment response from pre-treatment task-based fMRI and demonstrates a 26% improvement in performance in predicting response using augmented images. This improvement compares favorably to state-of-the-art augmentation methods for natural images. Through an ablative test, augmentation is also shown to substantively improve performance when applied before hyperparameter optimization. These results suggest the optimal order of operations and support the role of data augmentation method for improving predictive performance in tasks using fMRI.

1. INTRODUCTION

Neural networks have proved to be powerful modeling tools for many medical imaging problems, such as predicting neurological and psychiatric diagnoses and prognoses from brain MRI. However, the training of neural networks for these problems is frequently hindered by small dataset sizes, making it challenging to produce high-performing, generalizable models. Data augmentation, which synthesizes additional data samples from real data, has improved performance in many non-medical deep learning problems such as natural image classification. For example, recent methods such as AutoAugment[1] and Population-Based Augmentation[2] have improved classification error rate on highly studied datasets such as CIFAR, SVHN, and MNIST by up to 1.5% (a 12% relative improvement from previous state-of-the-art). On a reduced CIFAR dataset, the performance benefit of these AutoAugment was as high as 7%, highlighting the importance of data augmentation in cases of limited dataset size.

However, data augmentation techniques developed for natural images typically involve color and intensity transformations and geometric operations such as shearing, which may not be suitable for brain images because they introduce transformations that do not yield realistic brain appearance and morphology. In other words, these operations can produce implausible brain images. For brain MRI, one method for augmenting structural MRI (sMRI), involving independent components analysis (ICA) and random loading matrices, has shown to improve the accuracy of schizophrenia vs. healthy control classification by 5% (7-8% relative improvement using augmentation).[3,4] No such method has been developed and validated for functional MRI (fMRI).

The contributions of this work are as follows. 1) A novel, coregistration-based fMRI data augmentation method is proposed, which synthesizes new realistic raw fMRI images. 2) The performance benefit of this augmentation is demonstrated on an antidepressant response prediction task, where the goal is to produce a pre-treatment predictor of clinical response to a commonly used antidepressant, sertraline. Since individual antidepressant response is highly variable, improving the accuracy of such a predictor would help reduce morbidity in Major Depressive Disorder (MDD) by aiding clinicians in identifying MDD patients most likely to benefit from sertraline. 3) Additionally, this work provides evidence that augmentation not only improves overall model performance but also enables the identification of better models during model hyperparameter optimization.

Medical Imaging 2020: Image Processing, edited by Ivana Išgum, Bennett A. Landman,
Proc. of SPIE Vol. 11313, 113130T · © 2020 SPIE · CCC
code: 1605-7422/20/$21 · doi: 10.1117/12.2548630

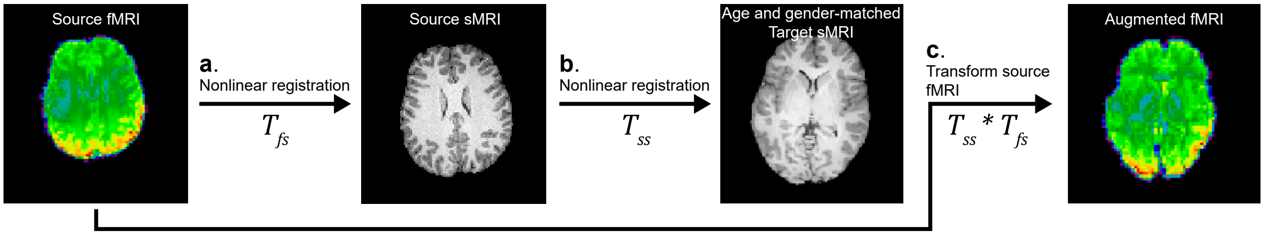

Figure 1. The proposed data augmentation method synthesizes a new fMRI image by performing a T1-based coregistration to another subject's brain in native space. **a.** The *source* fMRI mean frame is registered to the *source* sMRI. **b.** The *source* sMRI is registered to an age- and gender-matched *target* sMRI. **c.** The combination of these transformations is applied to transform the *source* fMRI into a synthetic fMRI in *target* space.

2. METHODS

2.1 Materials

Data from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) study,[5] a randomized controlled clinical trial, was used for the following experiments. This dataset contains 163 MDD subjects who underwent pre-treatment sMRI and task-based fMRI, then completed an 8-week treatment course with the antidepressant sertraline. The predictive task is to estimate the change in clinical severity between pre-treatment and week 8 of treatment from pre-treatment fMRI. This severity is measured by the Hamilton Rating Scale for Depression (HAMD) score. Demographics and pre-treatment clinical measurements, including psychiatric scales, comorbidities, and disease duration, are also added to the predictive models as covariates.

T1-weighted sMRI was acquired at 3T with the MPRAGE sequence; TE was 2.4 ms or 3.7 ms depending on study site, with dimensions of 256 × 256 × 176, and an isotropic voxel size of 1 mm. BOLD fMRI was acquired using GE-EPI, a TR of 200 ms, dimensions of 64 × 64 × 39, and voxel size of 3.2 × 3.2 × 3.1 mm for 8 minutes. During fMRI acquisition, subjects completed a block-design number-guessing task that stimulates reward processing circuitry known to be altered in depressed individuals.[6]

2.2 Data Augmentation

To generate anatomically-constrained synthetic fMRI images from real data, the proposed method employed a T1-based coregistration scheme to precisely resample a *source* subject's original fMRI signal onto the brain anatomy of a *target* subject in native space (**Fig. 1**). First, brain extraction was performed on the *source* and *target* subjects' **s**MRI using the ROBEX software. Brain extraction was then performed on the *source* subject's mean **f**MRI volume using a combination of FSL BET and AFNI 3dAutomask tools. In all cases, brains were manually inspected to confirm high quality brain extraction. Next (**1a**), the *source* fMRI mean volume was coregistered to the source sMRI using the `antsRegistrationSyNQuick` routine in ANTS, which performs a sequence of multi-scale rigid, affine, and nonlinear registration steps. Then (**1b**), the *source* sMRI was coregistered to the *target* sMRI using `antsRegistrationSyN` which accurately coregisters brain anatomy in sMRI across different subjects. Finally (**1c**), the transformations from steps **a**) and **b**) were combined and applied to the *source* fMRI to produce a new image with the *source*'s fMRI signal in the *target*'s brain space. This data augmentation effectively created *geometric* variation through both the nonlinear registration process and *intensity* variation inherently through the voxel interpolation during the warping transformation.

2.3 fMRI Preprocessing

The following preprocessing pipeline was applied to all original and augmented fMRI data: images were head-motion corrected through affine realignment of frames, brain-extracted as described in sect. 2.2 above, spatially normalized to the MNI152 EPI template, and smoothed with a 6 mm Gaussian kernel. Note that in contrast to the T1-based coregistration used during data augmentation, where the priority was to accurately warp a subject's brain into a different anatomy, a direct EPI-based registration was used for fMRI spatial normalization. Direct warping of individual fMRI images onto an EPI template has been shown to be more accurate for normalization to a template than cross-modal normalization as it accounts for magnetic inhomogeneities particular to EPI images.[7,8] The preprocessed fMRI images were fitted to subject-level generalized linear models (GLMs) in SPM12. The design matrix for the GLMs was defined as described in Greenberg et al.,[6] with regressors for each of the 3 conditions in the reward processing task. The fitted GLM coefficients for these regressors were projected back into voxel space to yield 3 contrast maps, i.e. spatial maps of BOLD response to each task condition. A study-specific brain parcellation was created from resting-state fMRI from all subjects in the EMBARC dataset using the spectral clustering method developed by Craddock et al.[9] This brain parcellation was used to compute the mean regional contrast values from each of the 3 contrast maps. These mean regional values are the input features for predictive model training, explained in the next section. The granularity of this parcellation (number of regions-of-interest [ROIs]) was optimized during the model selection process, during which 100-, 200-, and 400-ROI parcellations were tested.

2.4 Neural Network Construction, Model Search, and Validation

A feed-forward fully connected neural network was chosen as the predictive model. The model takes as input mean ROI values from each contrast map plus demographic and clinical covariates, and it predicts the 8-week change in the HAMD depression score. A loss function based on the coefficient of determination (R^2) was used:

$$L(\mathbf{y}, \hat{\mathbf{y}}) = 100(1 - f_{R^2}(\mathbf{y}, \hat{\mathbf{y}})) + \lambda$$

where $\mathbf{y}, \hat{\mathbf{y}}$ are the true and predicted HAMD scores and $f_{R^2}(.)$ computes the coefficient of determination for the set of points given through its arguments. The coefficient of 100 was chosen empirically to keep the magnitude of the loss roughly equal to that of the weight regularization term λ. Random search, a popular model selection method, was performed to optimize model hyperparameters, such as number of layers, neurons per layer, activation function, dropout rate, learning rate, and parcellation granularity. Three hundred (300) model configurations were randomly chosen over a predetermined hyperparameter search space (**Table 1**). To obtain an unbiased estimate of real-world performance, the models were evaluated using nested K-fold cross-validation with 5 inner folds and 3 outer folds. Within each outer fold, model performance was ranked by mean R^2 across the held-out partitions of the inner folds. The model with the highest R^2 was selected from each outer fold and test R^2 was measured on the held-out partition of the outer fold, not used for model training nor selection. The final model performance was the mean test performance over the 3 outer folds.

Table 1. Hyperparameter ranges used during random search.

Hyperparameter	Range searched
Model-level hyperparameters	
Number of fully-connected layers	1, 2, 3
Size of first layer	64, 96, 128, ..., 512
Weight regularization	L1, L2, L1 & L2
Nadam learning rate	0.0010, 0.0011, 0.0012, ..., 0.0030
Parcellation granularity (number of regions)	100, 200, 400
Layer-level hyperparameters	
Layer size taper rate*	0.5, 0.75
Batch normalization	Yes, No
Dropout rate	0.3, 0.4, 0.5, ..., 0.9
Activation	ReLU, Leaky ReLU, ELU, PReLU

*Taper rate determined the size of subsequent layers after the first fully-connected layer. Layer i has $n_i = t_i * n_{i-1}$ neurons, where t_i is the taper rate selected for that layer and n_{i-1} is the size of the previous layer.

Model searches were performed *without* data augmentation (which we denote as the search *Base*) and *with* the proposed data augmentation method (search *Aug*). To ensure fair comparisons, the 300 searched models, their initial weights, and the cross-validation splits were identical between the model searches. In search *Aug*, augmentation was applied to the training partition of each fold to increase sample size by a factor of 5. Specifically, each source subject in the training partition was augmented to 5 target subjects of the same gender and age decile. Target subjects were selected from other treatment groups without subject overlap in the EMBARC dataset to ensure that no target subjects came from the held-out partition. Demographics and clinical measurements were copied over to the 5 new synthetic samples without augmentation, so that the augmentation would only be synthesizing new fMRI data of a different but realistic brain shape, while assuming the remainder of the synthesized subject was constant.

2.5 Comparison with Affine Augmentation

A basic augmentation using affine co-registration was also performed for comparison. Specifically, the FLIRT tool from FSL was used to compute the register the source fMRI to the target sMRI.[10,11] As with the proposed, nonlinear augmentation method, each source subject was augmented to 5 gender- and age-matched target subjects. A third, identical model search (search *Aff*) was performed on this augmented data.

3. RESULTS

3.1 Data Augmentation

Synthesized contrast maps are compared to the original source fMRI in **Fig. 2**. While all three contrast maps have similar high and low intensity areas globally across the brain, locally the shape and location of the individual clusters do vary (*yellow circled*) which confirms that the augmentation procedure generated data samples with distinct spatial variations, yet with fidelity to the original contrast map.

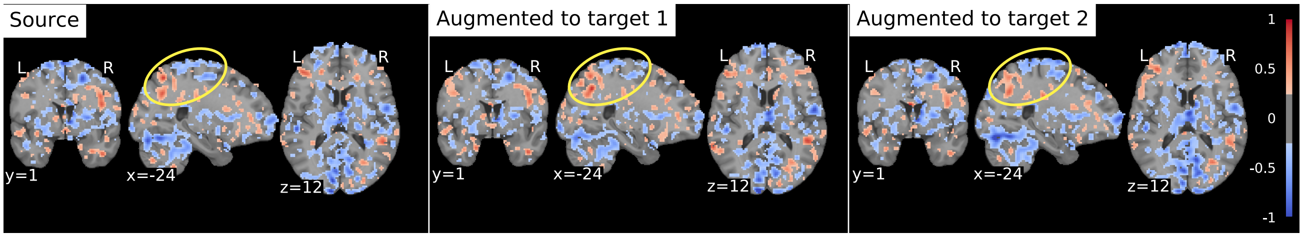

Figure 2. Example fMRI GLM contrast maps derived from a source fMRI image and two augmented fMRI images synthesized from the source. A region with similar distributions of high and low intensity areas but distinct local variations is circled in yellow. Contrast map values were thresholded at 0.25 for clarity.

3.2 Model Search Results

The top model from the non-augmented model search *Base*, referred to here as $Base_1$, achieved R^2 of 11.2% and RMSE of 6.57 in predicting HAMD change (**Table 2**). The top model from search *Aug* with the proposed augmentation, here called Aug_1, attained R^2 of 14.1% and RMSE of 6.46. This constitutes a substantial 26% relative increase in R^2 over $Base_1$. In comparison, the affine augmentation yielded a top model, Aff_1, with R^2 of 0.114 and RMSE of 6.531, a 2% relative increase in R^2 from $Base_1$. To ascertain the significance of this finding, these top models were retrained 100 times each with random weight initializations and R^2 and RMSE were compared with a two-tailed t-test. The differences in R^2 and RMSE between Aug_1 and $Base_1$ and between Aff_1 and $Base_1$ were both significant with $p = 0.001$. Comparisons of these performance gains to those in the literature are described in the discussion section below.

Table 2. Top model performance for each augmentation method.

Augmentation method	RMSE	R^2
Baseline (no augmentation)	6.57	0.112
Proposed (nonlinear)	**6.46**	**0.141**
Affine	6.53	0.114

3.3 Ablative Experiment: effect of data augmentation on model training

To test the impact of data augmentation in model training, Aug_1 was retrained without augmented data to create Aug'_1. Without the augmented data, performance decreased from R^2 of 14.1% and RMSE of 6.46 to R^2 of 10.7% and RMSE of 6.99. As a further test, the top 5 models from each of the 3 outer folds of search Aug ($Aug_1, Aug_2, ..., Aug_{15}$) were retrained without augmented data creating $Aug'_1, Aug'_2, ..., Aug'_{15}$. A pairwise two-tailed t-test between Aug_i and Aug'_i demonstrated *a significant decrease in performance when augmented data was removed*: R^2 decreased by $5.8 \pm 5.1\%$ ($p = 0.0006$) and RMSE increased by 0.21 ± 0.18 ($p = 0.0006$).

3.4 Additive Experiment: effect of data augmentation on model selection

In a reciprocal test of the impact of data augmentation in model selection, $Base_1$ was retrained with augmented data. This new model, $Base'_1$ did not exhibit increased performance with R^2 of 11.2% and RMSE of 6.74. This comparison was extended to the top 5 models from each of the 3 outer folds of search $Base$ ($Base_1, Base_2, ..., Base_{15}$), which were retrained with augmented data creating $Base'_1, Base'_2, ..., Base'_{15}$. A pairwise two-tailed t-test between $Base_i$ and $Base'_i$ revealed a non-significant improvement: R^2 increased by $1.5 \pm 4.4\%$ ($p = 0.209$) and RMSE decreased by 0.05 ± 0.16 ($p = 0.214$).

4. DISCUSSION AND CONCLUSION

This work introduces a data augmentation method for synthesizing fMRI images with realistic brain morphology and demonstrates its performance benefit in a predictive task. The overall best performance in predicting antidepressant response was achieved by performing augmentation before model search. The best model using this approach, Aug_1, outperformed the best model from a search conducted without augmentation, $Base_1$, by 26% in R^2. Additionally, using affine transformations to augment the data did not substantially improve performance, indicating that the nonlinear co-registration performed in the proposed method is integral to its effectiveness. The rudimentary geometric variations and minimal voxel intensity noise generated by affine transformations were likely canceled out once the augmented fMRI images were all spatially normalized to a common template. However, the more geometrically complex transformations and voxel interpolation effects introduced by the proposed nonlinear augmentation method were not fully canceled out during spatial normalization. The resulting preprocessed fMRI data shows distinct variations in the derived contrast maps (**Fig. 2**).

The effectiveness of this augmentation method compares favorably with the 12% relative improvement achieved by state-of-the-art augmentation methods for natural images[1,2] and the 7-8% relative improvement achieved by a previous sMRI augmentation method.[3,4] Additionally, this method was shown to provide the most benefit when used not only for model training, but also throughout the model search process. In the ablative experiment, models selected from search Aug performed significantly worse when retrained without augmented data. In fact, the previous best model Aug'_1 performed worse than $Base_1$, suggesting that the high-performing models found in search Aug would have been missed by search $Base$. These observations strongly indicate that data augmentation should be performed prior to model hyperparameter optimization. Conversely, the additive experiment showed a lesser performance benefit when augmented data was introduced after the model search. This suggests that search $Base$ identified less statistically powerful models that could not increase performance when data was augmented after the search.

While these results were limited to one task-based fMRI dataset, additional work will demonstrate generalization to additional datasets and resting-state fMRI. Another possible limitation arose from the T1-based coregistration employed in the augmentation. The source fMRI-to-source sMRI cross-modal registration may be imprecise due to EPI-specific non-linearities, causing the source fMRI to not be exactly registered to the target brain. Future work may test a direct EPI-based coregistration to the target fMRI. Finally, future experiments will test more extensive augmentation such as to 10-20x the original dataset size rather than 5x. Despite these limitations, the current findings show that the proposed fMRI augmentation can already significantly improve deep learning performance on neuroimaging predictive tasks.

In conclusion, this work proposes a novel, coregistration-based fMRI data augmentation method to synthesize realistic fMRI images that requires no new expensive fMRI acquisition. The method demonstrates improved performance in a challenging prediction task of antidepressant response prediction. This work also provides evidence that augmentation should precede hyperparameter optimization and that augmentation not only improves overall model performance but also the identification of better models during model hyperparameter optimization. We look forward to extending this promising approach to further increase its benefits.

REFERENCES

[1] Cubuk, E. D., Zoph, B., Mane, D., Vasudevan, V., and Le V, Q., "AutoAugment: Learning Augmentation Policies from Data," in [*Conference on Computer Vision and Pattern Recognition (CVPR)*], (2019).

[2] Ho, D., Liang, E., Stoica, I., Abbeel, P., and Chen, X., "Population Based Augmentation: Efficient Learning of Augmentation Policy Schedules," in [*International Conference on Machine Learning (ICML)*], (2019).

[3] Ulloa, A., Plis, S., Erhardt, E., and Calhoun, V., "Synthetic structural magnetic resonance image generator improves deep learning prediction of schizophrenia," in [*2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*], 1–6, IEEE (2015).

[4] Castro, E., Ulloa, A., Plis, S. M., Turner, J. A., and Calhoun, V. D., "Generation of synthetic structural magnetic resonance images for deep learning pre-training," in [*2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*], 1057–1060, IEEE (2015).

[5] Trivedi, M. H., McGrath, P. J., Fava, M., Parsey, R. V., Kurian, B. T., Phillips, M. L., Oquendo, M. A., Bruder, G., Pizzagalli, D., Toups, M., Cooper, C., Adams, P., Weyandt, S., Morris, D. W., Grannemann, B. D., Ogden, R. T., Buckner, R., McInnis, M., Kraemer, H. C., Petkova, E., Carmody, T. J., and Weissman, M. M., "Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design," *Journal of Psychiatric Research* **78**, 11–23 (2016).

[6] Greenberg, T., Chase, H. W., Almeida, J. R., Stiffler, R., Zevallos, C. R., Aslam, H. A., Deckersbach, T., Weyandt, S., Cooper, C., Toups, M., Carmody, T., Kurian, B., Peltier, S., Adams, P., McInnis, M. G., Oquendo, M. A., McGrath, P. J., Fava, M., Weissman, M., Parsey, R., Trivedi, M. H., and Phillips, M. L., "Moderation of the Relationship Between Reward Expectancy and Prediction Error-Related Ventral Striatal Reactivity by Anhedonia in Unmedicated Major Depressive Disorder: Findings From the EMBARC Study," *The American Journal of Psychiatry* **172**(9), 881–891 (2015).

[7] Dohmatob, E., Varoquaux, G., and Thirion, B., "Inter-subject Registration of Functional Images: Do We Need Anatomical Images?," *Frontiers in Neuroscience* **12** (2018).

[8] Calhoun, V. D., Wager, T. D., Krishnan, A., Rosch, K. S., Seymour, K. E., Nebel, M. B., Mostofsky, S. H., Nyalakanai, P., and Kiehl, K., "The impact of T1 versus EPI spatial normalization templates for fMRI data analyses," *Human Brain Mapping* **38**(11), 5331–5342 (2017).

[9] Craddock, R. C., James, G. A., Holtzheimer, P. E., Hu, X. P., and Mayberg, H. S., "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Human Brain Mapping* **33**(8), 1914–1928 (2012).

[10] Jenkinson, M. and Smith, S., "A global optimisation method for robust affine registration of brain images," *Medical image analysis* **5**(2), 143–156 (2001).

[11] Jenkinson, M., Bannister, P., Brady, M., and Smith, S., "Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images," *NeuroImage* **17**(2), 825–841 (2002).