Detection and removal of biases from brain MR images using adversarial architectures

Sahitya Mantravadi
Stanford University
Institute for Computational and Mathematical Engineering

Introduction

To practically combine MRI datasets from different sites, we must be able to remove biases in each image that point to the site at which that MR image was taken. The biases in this case are any markers or difference between MR images from different sites. To achieve this, we outline the architecture that will preserve information in each image and make accurate predictions while mitigating the presence of confounds. Using the ABIDE dataset, we aim to develop an MRI debiasing architecture that will minimize the accuracy of site prediction and maximize the accuracy of quality prediction. See Figure 1 for an overview of the model architecture

Data

We apply methods to the Autism Brain Imaging Data Exchange dataset, which aggregates smaller sets of data from 17 different sites. Each MR image is represented as a three-dimensional matrix, capturing three views of the brain: axial, coronal, and sagittal. In total, the dataset contains 1103 MRI. This was randomly split into 803 training examples, 200 validation examples, and 100 test examples.
Because the architecture encapsulates two discriminators, each data point is associated with two labels: a quality score label (0 or 1) and a site label (0 to 16). Both labels are victims to class imbalance. See Table 1 for class breakdown per label. We attempt to mitigate this by using rejection resampling for quality score labels in each batch of training data.

Site	Good Quality	Bad Quality	Total
0	35	3	38
1	23	4	27
2	44	11	55
3	48	16	64
4	45	12	57
5	158	26	184
6	27	1	28
7	19	17	36
8	46	11	57
9	22	8	30
10	13	23	36
11	0	40	40
12	40	9	49
13	72	28	100
14	35	110	145
15	96	5	101
16	19	37	56
Total	742	361	1103

Table 1. Number of examples per class for each label

Defining Losses

To calculate loss for each network, we examine the logits produced by each of the discriminator networks. We compare the logits to the true labels by using the cross entropy loss.

Site discriminator loss – We use cross-entropy loss for 17 classes.

$$L^s = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{16} y_{ij}^s \log p_{ij}^s$$

Quality discriminator loss – We use cross-entropy loss for 2 classes.

$$L^q = -\frac{1}{n}\sum_{i=1}^{n} y_i^q \log p_i^q + (1-y_i^q)\log(1-p_i^q)$$

De-biasing generator loss – We use the weighted difference of the two discriminator losses. Thus, by minimizing the loss of the generator, we decrease the quality discriminator loss and increase the site discriminator loss.

$$L^G = W_q L^q - W_s L^s$$

Tuning the weights on each loss tunes the accuracy/debiasing tradeoff. With a reduction in site information in each de-biased image, we expect a marginal decrease in quality accuracy.

Architectures + Results

The final architecture of the model is described below. Utilizing skip connections within the generator made a large difference in results.

Discriminators:

1. 3D convolution with 32 kernels of size 4 and stride 1
2. Max pooling with window size of 2 and stride of 2
3. Leaky ReLU
4. 3D convolution with 16 kernels of size 4 and stride 1
5. Max pooling with window size of 2 and stride of 22
6. Batch normalization
7. Leaky ReLU
8. Dropout
9. 3D convolution with 8 kernels of size 4 and stride 1
10. Max pooling with window size of 2 and stride of 2
11. Batch normalization
12. Leaky ReLU
13. Dropout with
14. 3D convolution with 1 kernel of size 4 and stride 1
15. Max pooling with window size of 2 and stride of 2
16. Fully connected layer with 100 units
17. Leaky ReLU
18. Dropout
19. Fully connected layer with 17 units (2 units) for site discriminator (quality discriminator)

Generator:

1. 3D convolution with 16 kernels of size 4 and stride 2
2. Leaky ReLU
3. 3D convolution with 32 kernels of size 4 and stride 2
4. Batch normalization
5. Leaky ReLU
6. Dropout
7. 3D transposed convolution with 16 kernels of size 4 and stride 2
8. Batch normalization
9. Tanh
10. Resizing/padding to be of the same size as the output of Layer 2
11. Concatenation with Layer 2
12. 3D transpose convolution with 4 kernels of size 4 and stride 2
13. Batch normalization
14. Leaky ReLU
15. Dropout
16. 3D convolution with 1 kernel of size 4 and stride 1

A natural control for our framework is the same setup as Figure 1, but with the generator replaced by the identity function. This is mathematically equivalent to training two separate convolutional neural networks, one to predict site and another to predict quality. Utilizing the discriminator architecture detailed earlier, we achieved 85% accuracy for quality prediction (due to noisy labels) and 96% accuracy for site prediction. Figures 2 and 3 show an MR image before and after de-biasing.
After de-biasing, the quality discriminator achieved 76% accuracy on the test set, and the site discriminator achieved 12% accuracy on the test set.

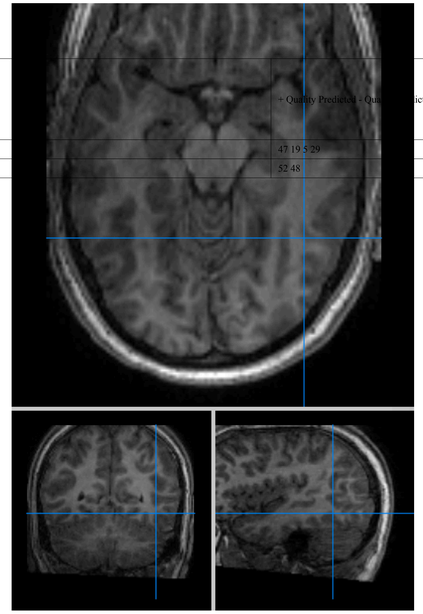

Figure 2. Original MR image

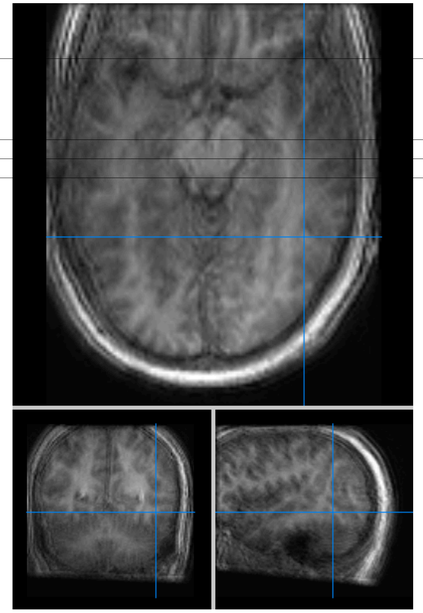

Figure 3. De-biased MR image

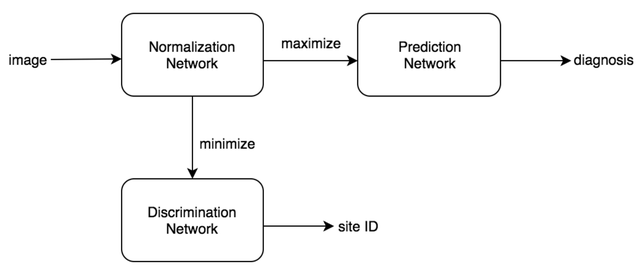

Figure 1. Overview of model architecture

Acknowledgements + Selected References

I would like to thank Chris Gorgolewski, a postdoctoral researcher with the Stanford Center for Reproducible Neuroscience for his help in outlining the project, providing access to data, and mentorship and advice throughout the project.
I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. *CoRR*, abs/1406.2661, 2014.
B. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. 2018. CoRR, abs/1801.07593.
P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.
M.MirzaandS.Osindero.Conditionalgenerativeadversarial nets. *CoRR*, abs/1411.1784, 2014.
O. Esteban, D. Birman, M. Schaer, O. Koyejo, R. Poldrack, and K. Gorgolewski. Mriqc: Advancing the automatic predic- tion of image quality in mri from unseen sites. 2017. PLOS ONE 12(9):e0184661.10.1371/journal.pone.0184661.