Detection and removal of biases from brain MRI using adversarial architectures

Sahitya Mantravadi
Stanford University
Institute for Computational and Mathematical Engineering
smantra@stanford.edu

Chris Gorgolewski
Stanford University
Center for Reproducible Neuroscience
chrisgor@stanford.edu

Abstract

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 MRI 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.

We then detail the methods and architectures tested in the development of the debiasing network. We discuss pitfalls and issues encountered. Lastly, we discuss related work and the extensibility of this architecture for other use cases.

1. Introduction

Neuroscience researchers can now leverage advances in the accuracy and explainability of convolutional neural networks to draw conclusions from MRI and fMRI data. However, reproducible neuroscience is difficult to achieve without large datasets. Hand-collected and hand-labelled MRI data can be difficult to aggregate across medical sites, especially when scanners have disparate settings and MR images have different sizes and resolutions. If not properly dealt with, differences brought about by site and scanner variation may be falsely attributed to biological differences. In order to mitigate these site effects, the ultimate goal is to remove site effects from the data, effectively "de-biasing" each image.

We develop and apply adversarial methods to a set of brain MRI data. For each image, we have a quality score label given to the image by a rater. Intuitively, preserving this quality score and retaining the ability to predict this quality score while de-biasing data allows us to preserve the quality of and important information from each image while removing site bias. To do this, we evaluate several adversarial network architectures in the context of removal of such biases from MR images. On a high level, we create an adversarial network architecture that has three components as shown in Figure 1 (building on the structure of a conditional generative adversarial network

1. a discriminator to predict the site of an input image
2. a discriminator to predict the quality of an input image
3. a de-biasing component (here, the generator) that takes the input image and outputs a modified image. This component minimizes the accuracy of (1) and maximizes the accuracy of (2)

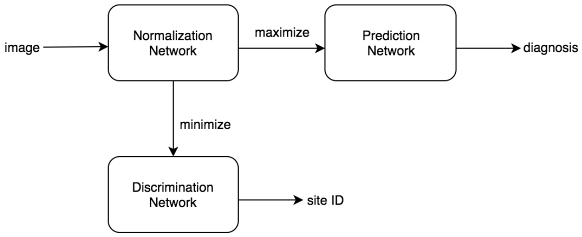

Figure 1. The three components of the adversarial architecture.

We will describe the architecture of each component as well as specific design choices that were made.

2. Related Work

3. Data

We apply methods to the Autism Brain Imaging Data Exchange (ABIDE I) dataset, which aggregates smaller sets of data from 17 different sites, as described in detail and utilized in

Each MRI is represented as a three-dimensional matrix, which can be visualized with the MRI software Mango

Because this dataset consists of 17 sets of images from different sites, images from a particular site can have different sizes and resolutions than images from other sites. The MRI have all been standardized to be of size 106 × 128 × 110. The output images from the debiasing component of the architecture are of this standard size as well. No data augmentation or other preprocessing steps were taken.

In total, the dataset contains 1103 MRI. This was randomly split into 803 training examples, 200 validation examples, and 100 test examples. Due to RAM constraints, we were limited to a batch size of 4 to train. This means each epoch over the training data is about 200 training steps.

3.1. Labels

Because the architecture encapsulates two discriminators, each data point is associated with two labels: a quality score label and a site label.

Each site label is simply which site the brain MRI is from, one of the following 17 sites (corresponding class as an integer):

- California Institute of Technology (0)
- Carnegie Mellon University (1)
- Kennedy Krieger Institute (2)
- Ludwig Maximilians University Munich (4)
- NYU Langone Medical Center (5)
- Olin, Institute of Living at Hartford Hospital (7)
- Oregon Health and Science University (6)
- San Diego State University (10)
- Social Brain Lab, BCN NIC UMC Groningen and Netherlands Institute for Neurosciences (9)
- Stanford University (11)
- Trinity Centre for Health Sciences (12)
- University of California, Los Angeles (13)
- University of Leuven (3)
- University of Michigan (14)
- University of Pittsburgh School of Medicine (8)
- University of Utah School of Medicine (15)
- Yale Child Study Center (16).

Thus, the site label is for one of 17 classes.

Each image has at least one but up to 3 quality scores (-1, 0, or +1), given by a human rating the quality of the MRI. To build the quality score label, we averaged the human ratings and assigned a quality score of 1 if the average rating was greater than 0; the MRI was assigned a quality score of 0 is the average rating was less than or equal to 0. Thus, the quality score label is for one of 2 classes.

Site	+ Quality	- Quality	Total MRIs
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
-	742	361	1103

Table 1. Examples per site label class and quality label class

3.2. Class Imbalance

In the ABIDE II dataset, both the site and quality score labels are victims of class imbalance, which can heavily impact the performance of any algorithm. Table 1 shows the number of examples in each class for each label.

We can see here that about 70% of the data has a ’good’ quality label of 1, so only 30% has a ’bad’ quality label of 0. It is also important to note that several sites (5, 14, 15, 13) have significantly more examples than other sites. In addition, some of the sites (11, 14) are highly correlated with poor quality MR images. while other sites (0, 1, 6, 15) are highly correlated with good quality MR images.

In our data pipeline, each training batch is rejection resampled to contain 50% good quality and 50% bad quality MR images. Rejection resampling for both labels at the same time proved inefficient, but a better solution to this dual class imbalance should be developed. In our results, we note that balancing just the quality label for each batch mitigates some of the class imbalance for the site labels as well.

4. Algorithms, Architectures, and Model

4.1. Layers

Because we have three-dimensional images, we utilize three-dimensional convolutions in each architecture. Each kernel for the three dimensional convultional layer spans 4 pixels.

4.2. Algorithms

4.3. Loss

5. Methods

We update the

6. Experiments and Results

7. Conclusion

References