CM3203 One Semester Individual Project

Initial Plan

Comparing Interpretability Metrics for Diagnostic Classification of Neuropsychiatric Disorders Using Clinical MRI Data

Author: Elise Bailey

Supervisor: Matthias Treder

February 3rd 2020

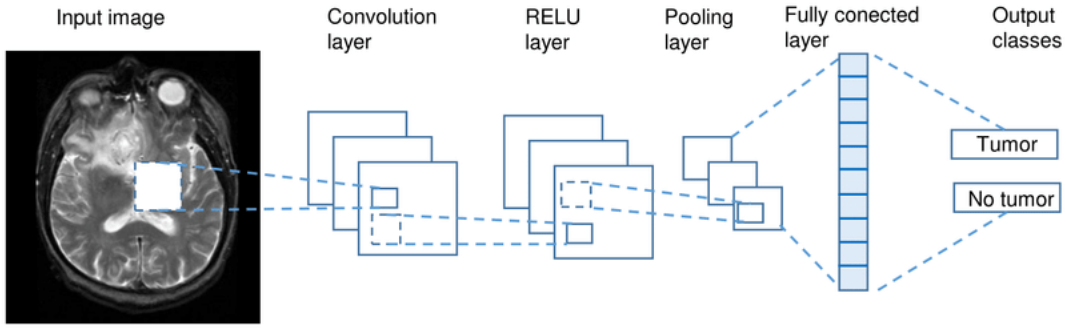
Project Description

Neuropsychiatric disorders are a group of medical conditions that involve both neurology and psychiatry. This means they have a physical component relating to the nervous system as well as a mental component that causes emotional distress and abnormal behaviour. The disorders that the data I will be using comprises of are schizophrenia, posttraumatic stress disorder (PTSD) and Parkinson's disease (PD). These disorders develop as a result of complex genetics combined with environmental impacts throughout the patient's lifespan. As a result, the physical anatomy of the brain is altered. For example, in patients with PD loss of grey matter in the hypothalamus is often observed which causes a reduction in the levels of dopamine, serotonin, melanin and hypocretin.[1] Dopamine is responsible for relaying messages that plan and control body movement, hence a reduction causes the signature symptoms of PD; a tremor, slow movement and stiffness.[2]

Clinical diagnosis of these disorders is still a challenge as they are very complex and largely still not understood. However, the physical changes in the brain mean that machine learning can be used to differentiate patients using MRI scans. An MRI is a 3D imaging method that uses strong magnetic waves to produce detailed images of the organs and tissues within the body. MRIs are therefore very suited to detecting changes in volume of grey matter (a type of tissue in the brain), a key sign of a neuropsychiatric disorder, as well as other neuroanatomical changes that are characteristic of such disorders.

For the initial part of my project I will be using deep learning on clinical MRI scans to determine whether the patient is healthy or has a neuropsychiatric disorder. Deep learning is a branch of machine learning that was actually inspired by the human brain. It creates artificial neural networks that are able to learn by experience and acquire skills without human

involvement from a large amount of data. More specifically, I will be implementing a convolutional neural network (CNN) which is a class of deep learning that is capable of image classification. In the context of deep learning, image classification is the process of taking an image as an input and outputting a class that it most closely belongs to. In this project the input will be an MRI scan and the output will indicate that the patient to which it belongs is either healthy, a schizophrenia patient, a PD patient or a PTSD patient. A CNN is comprised of multiple layers: convolutional layers, ReLU layers, pooling layers and a fully connected layer. In simple terms, the convolution, ReLU and pooling layers break up the image into features and analyses them. The fully connected layer then takes the outputs of these layers as an input and uses them to make a classification decision.



Ker, J., Wang, L., Rao, J. and Lim, T. (2017).

Figure 1: An illustration of how a CNN can be used to detect a brain tumour in an MRI. In: Deep Learning Applications in Medical Image Analysis. IEEE Access, 6, pp.9375-9389

This method of deep learning has already been used to extract patterns from neuroimaging data, most notably in a project to distinguish Alzheimer's MRIs from healthy control MRIs.[3] The aims of this project and the data it used are closely linked to the initial aim of my project and the data I will be using so by adapting this and using available machine learning libraries creating a CNN for this task should be achievable in a relatively small timeframe.

The second part of my project, comparing interpretability metrics, is more challenging because CNNs are notoriously uninterpretable.[4] This is because they consist of many hidden layers, all of which make important but complex computational decisions, the reasonings behind which are often difficult to understand. This lack of transparency means the resulting computation can be difficult to interpret and can have counter-intuitive properties.[5] Transparency and interpretation are particularly important in the context of neuropsychiatric diagnosis because it is important to understand why false-positives and false-negatives may occur. When diagnostic errors happen due to human error, accountability and liability are investigated to find out what caused the error and how to prevent a similar situation occurring again. Therefore, to gain patient and provider trust, a deep learning model should be as transparent as possible.[6] For this to be the case, it must be clear how the model uses the input

data and how it makes decisions. I will therefore be analysing and evaluating a range of interpretability toolboxes that aim to explain deep learning models. As I will be developing the CNN using Keras, a Python deep learning library, the toolboxes I will be evaluating will also use Python. Namely, I will start by evaluating yellowbrick, ELI5, LIME and MLxtend.[7] One of the biggest challenges will be finding a fair, quantitative way to measure just how interpretable these toolboxes make the models. The overall goal of my project is therefore to objectively and fairly evaluate how well these different toolboxes produce consistent, interpretable and useful information on the diagnostic classification of neuropsychiatric disorders using a convolutional neural network.

Ethics

I will be using clinical MRI data provided by the Shared Roots project at Stellenbosch University. The study is approved by the Health Research Ethics Committee at Stellenbosch University and I will attach the ethics form to my final report. I will also forward these documents to COMSC Ethics to ensure no further steps need to be taken before I start working with the data.

Project Aims and Objectives

Aims

- Develop a CNN to classify MRI scans of neuropsychiatric disorders
- Evaluate the interpretability of a range of Python libraries for interpreting deep learning models

Objectives

- Use Keras to implement a CNN that takes clinical MRIs as input and classifies them as one of; healthy, a schizophrenia patient, a PD patient or a PTSD patient.
- Train the CNN using the Share Roots data
- Experiment with a range of CNN architectures to determine which gives the most consistent results
- Analyse the different capabilities and features of a range of interpretability toolboxes
- Use the toolboxes to produce interpretable data about the CNNs decisions and results, with the aim to increase transparency
- Determine a method of quantitative evaluation of how interpretable and useful the data produced by these models is, and therefore by how much they increase the transparency of the CNN
- Draw conclusions about the limits of CNN interpretability and whether deep learning models can truly be transparent in the context of clinical MRI classification

Work Plan

Supervisor Meetings

There will be weekly group meetings with my supervisor, Matthias Treder, and other students working on machine learning projects where we will discuss what we've been doing and any problems that we have encountered. I will also have one-on-one review meetings every 3 weeks, or as needed, with my supervisor.

Weekly Plan

I have outlined the key tasks that I aim to complete by the end of each week however it is worth pointing out that it is likely to be a much more iterative than linear process so I may revisit previous tasks in later weeks. Additionally, it is likely that I will need to factor in unseen challenges and issues so I will need to be very flexible with my time and my goals for each week.

Week 1 27/01 – 02/02

- Write initial plan
- Ensure all software is downloaded
- Research CNNs

Week 2 03/02 – 09/02

- Research CNNs, classification algorithms and interpretation toolboxes
- Complete tutorials on Keras and implementing neural networks

Week 3 10/02 – 16/02

- Program a CNN to classify neuropsychiatric disorders

Week 4 17/02 – 23/02

- Program a CNN to classify neuropsychiatric disorders
- Start training the CNN using MRI scans

Week 5 24/02 – 01/03

- Further training and potential changes to the CNN
- Experiment with different CNN architectures

Week 6 02/03 – 08/03

- Experiment with different CNN architectures and evaluate their results

Week 7 09/03 – 15/03

- Examine the available Python interpretation toolboxes to decide which ones to evaluate
- Determine a quantitative way to evaluate the interpretability toolboxes

Week 8 16/03 – 22/03

- Use the interpretability toolboxes to visualise the CNN model's decisions and results

Week 9 23/03 – 29/03

- Use the interpretability toolboxes to visualise the CNN model's decisions and results

Easter 30/3 – 19/04

- Evaluate how good the different toolboxes are at increasing the transparency of the CNN model's decisions and results
- Write up approach and background for final report

Week 10 20/04 – 26/04

- Write up results and findings in final report

Week 11 27/04 – 03/05

- Write up results and findings in final report

Week 12 04/05 – 07/05

- Finalise and hand in report

References

[1] Prakash, K., Bannur, B., Chavan, M., Saniya, K., Sailesh, K. and Rajagopalan, A. (2016). Neuroanatomical changes in Parkinson′s disease in relation to cognition: An update. *Journal of Advanced Pharmaceutical Technology & Research*, [online] 7(4), p.123. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5052937/ [Accessed 30 Jan. 2020].

[2] Spine, M. (2018). *Parkinson's Disease (PD) Mayfield Brain & Spine Cincinnati, Ohio*. [online] Mayfieldclinic.com. Available at: https://mayfieldclinic.com/pe-pd.htm [Accessed 30 Jan. 2020].

[3] Sarraf, S., DeSouza, D., Anderson, J. and Tofighi, G. (2017). DeepAD: Alzheimer's Disease Classification via Deep Convolutional Neural Networks using MRI and fMRI. [online] Available at: https://www.biorxiv.org/content/10.1101/070441v4.full [Accessed 30 Jan. 2020].

[4] Zhang, Q., Wu, Y. and Zhu, S. (2018). Interpretable Convolutional Neural Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. [online] Available at: https://arxiv.org/pdf/1901.02413.pdf [Accessed 30 Jan. 2020].

[5] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R. (2014). *Intriguing properties of neural networks*. [online] arXiv.org. Available at: https://arxiv.org/abs/1312.6199 [Accessed 30 Jan. 2020].

[6] Geis, J., Brady, A., Wu, C., Spencer, J., Ranschaert, E., Jaremko, J., Langer, S., Kitts, A., Birch, J., Shields, W., van den Hoven van Genderen, R., Kotter, E., Gichoya, J., Cook, T., Morgan, M., Tang, A., Safdar, N. and Kohli, M. (2019). Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement. *Insights into Imaging*, [online] 10(1). Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6768929/ [Accessed 30 Jan. 2020].

[7] Vickery, R. (2019). *Python Libraries for Interpretable Machine Learning*. [online] Medium. Available at: https://towardsdatascience.com/python-libraries-for-interpretable-machine-learning-c476a08ed2c7 [Accessed 30 Jan. 2020].