2021-4

An Analysis of the Interpretability of Neural Networks trained on Magnetic Resonance Imaging for Stroke Outcome Prediction

Esra Zihni
Technological University Dublin, esra.zihni@tudublin.ie

John D. Kelleher
Technological University Dublin, john.d.kelleher@tudublin.ie

Bryony McGarry
Technological University Dublin, bryony.mcgarry@tudublin.ie

Recommended Citation
Zihni E., McGarry B.L, & Kelleher JD. (2021). An analysis of the interpretability of neural networks trained on magnetic resonance imaging for stroke outcome prediction. *Proc. Intl. Soc. Mag. Reson. Med*, vol. 29, pg. 3503. doi:10.21427/dhbt-q252

Funder: Precise4Q

An Analysis of the Interpretability of Neural Networks trained on Magnetic Resonance Imaging for Stroke Outcome Prediction

Esra Zihni[1], Bryony McGarry[1,2], and John D. Kelleher[1,3]
[1]PRECISE4Q, Predictive Modelling in Stroke, Information Communications and Entertainment Institute, Technological University Dublin, Dublin, Ireland, [2]School of Psychological Science, University of Bristol, Bristol, United Kingdom, [3]ADAPT Research Centre, Technological University Dublin, Dublin, Ireland

Synopsis

Applying deep learning models to MRI scans of acute stroke patients to extract features that are indicative of short-term outcome could assist a clinician's treatment decisions. Deep learning models are usually accurate but are not easily interpretable. Here, we trained a convolutional neural network on ADC maps from hyperacute ischaemic stroke patients for prediction of short-term functional outcome and used an interpretability technique to highlight regions in the ADC maps that were most important in the prediction of a bad outcome. Although highly accurate, the model's predictions were not based on aspects of the ADC maps related to stroke pathophysiology.

Introduction

Multiparametric MRI can assist the clinician in treatment decisions of acute ischaemic stroke patients.[1] Diffusion-weighted imaging (DWI) and ADC maps, reveal cytotoxic oedema, enabling diagnosis of ischaemia and, in combination with other MR parameters help evaluate tissue status.[2] In recent years, deep learning models have been effectively applied to medical image data.[3] Based on the rich pathophysiological information in ADC images of acute stroke patients, deep learning applied to ADC maps may yield strong predictive power. Convolutional neural networks (CNNs) are a form of deep learning designed specifically to process image data. A distinctive characteristic of CNNs is their ability to identify relevant local visual features irrespective of where they occur in the image.[4] CNNs have been used to predict the functional outcome of ischaemic stroke patients based on brain imaging.[5-7] However, a criticism of neural networks is that they lack transparency[8] meaning it is unclear what information in the MRI scan the CNN uses to make a prediction. Recently, several methods for interpreting neural networks have been developed.[9] Here, we applied an attention-based method to examine the decisions of a CNN trained to predict short-term functional outcome for hyperacute ischaemic stroke patients based on ADC maps. Our results indicate that although the CNN was accurate on the task, its decision was not based on biologically relevant information.

Methods

We used the ADC maps, lesion masks and modified Ranking Scale (mRS) scores from 40 hyperacute ischaemic stroke patients (mean onset time = 160 (72) minutes), from the ISLES (Ischemic Stroke Lesion Segmentation) 2017 challenge training dataset.[10,11] To standardize the image size, ADC maps were downscaled to 128 x 128 x 19 voxels using cubic spline interpolation. ADC maps were also normalized to have voxel intensity values with zero mean and standard deviation of one. The 90 days mRS scores were dichotomized where a score of 0-2 indicates good outcome (negative class) and 3-6 indicates bad outcome (positive class), resulting in 9 positive and 31 negative instances. We modelled 3D volumes of the ADC maps to predict bad outcome at 90 days, using a 3D-CNN with two convolutional and max pooling layers followed by a dense layer. For regularization we used l2 normalization on each layer and dropout on the dense layer. We split the data into training and validation sets with a 4:1 ratio while preserving class percentages in each set. We fine-tuned hyperparameters on the training set using 5-fold cross-validation with grid search, and area under the receiver operator characteristics curve (AUROC) as the evaluation metric. The final model hyperparameters are given in Figure 1. We evaluated the final model on the validation set using AUROC. We used gradient-based class activation mapping (grad-CAM)[12] to provide visual explanations on the final model's decisions, which provides localization of regions of interest on the input image that leads to the prediction of a target class. In our case, the target is the positive class representing a patient with a bad outcome. We visualized examples from training and validation sets and qualitatively compare images to investigate common patterns.

Results

The final model performed well both on the training and validation sets with an AUROC of 0.99 and 0.92 respectively, showing that CNNs applied to ADC maps of hyperacute ischaemic stroke patients can predict short-term functional outcome with high accuracy. Visual explanations of the model's decision in terms of a bad outcome are presented as heatmaps in Figures 2 and 3. Initial visual inspection shows that the most highlighted areas over all of them were the external boundaries with a focus on the front of the brain.

Discussion

The CAMs showed that the model did not focus on the visible ischaemic regions in the ADC maps, but consistently focused on the boundaries of the brain. This finding suggests the model's predictions were likely based on MR artifacts rather than pathophysiological information represented in the ischaemic regions of ADC maps. For example, movement artifacts, which are more likely to occur for extremely unwell patients. Eddy currents caused by changing magnetic fields during image acquisition are also particularly problematic in diffusion imaging and computed maps of diffusion parameters such as ADC.[13] Hence, we hypothesize that the model's decisions were more likely based on abnormalities during image acquisition rather than biological indications of stroke severity. These results highlight the issue that a high performing model is not necessarily a reliable model.

Conclusion

Application of CNNs to MRI has the potential to inform treatment decisions in acute ischaemic stroke, but their integration into the clinical setting will require a robust understanding of model decisions. This work highlighted that, contrary to the assumption that the predictions are based on biologically relevant information inherent in the image, they could be based on factors related to the MR acquisition. Understanding why a neural network behaves a certain way can provide insight into the model's weaknesses, which in turn may be improved through domain knowledge and modification of methodology. Future work will involve applying the same modelling and interpretability techniques on only the ADC defined ischaemic region, to improve chances of the network identifying biologically relevant information for prediction of short-term outcome.

Acknowledgements

This research was supported by the PRECISE4Q project, funded through the European Union's Horizon 2020 research and innovation program under grant agreement No. 777107, and the ADAPT Research Centre, funded by Science Foundation Ireland (Grant 13/RC/2106) and is co-funded by the European Regional Development fund.

References

1. Wintermark M, Albers GW, Alexandrov AV, et al. Acute stroke imaging research roadmap. AJNR Am J Neuroradiol. 2008;29(5):e23-e30. doi:10.1161/STROKEAHA.107.512319

2. Kauppinen RA. Multiparametric Magnetic Resonance Imaging of acute experimental brain ischaemia. Prog Nucl Magn Reson Spectrosc. 2014;80:12-25. doi:10.1016/j.pnmrs.2014.05.002

3. Yala A, Lehman C, Schuster T, Portnoi T, Barzilay R. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology. 2019;292(1):60-66. doi:10.1148/radiol.2019182716

4. Kelleher, J. D. (2019). Deep Learning. MIT Press.

5. Hilbert A, Ramos LA, van Os HJA, et al. Data-efficient deep learning of radiological image data for outcome prediction after endovascular treatment of patients with acute ischemic stroke. Comput Biol Med. 2019;115:103516. doi:10.1016/j.compbiomed.2019.103516

6. Bacchi S, Zerner T, Oakden-Rayner L, Kleinig T, Patel S, Jannes J. Deep Learning in the Prediction of Ischaemic Stroke Thrombolysis Functional Outcomes. Acad Radiol. 2020;27(2):e19-e23. doi:10.1016/j.acra.2019.03.015

7. Zihni E, Madai V, Khalil A, et al. Multimodal Fusion Strategies for Outcome Prediction in Stroke. In: Proceedings of the 13th International Joint Conference on Biomedical Engineering Systems and Technologies. SCITEPRESS - Science and Technology Publications; 2020:421-428. doi:10.5220/0008957304210428

8. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). IEEE Access. 2018;6:52138-52160. doi:10.1109/ACCESS.2018.2870052

9. Montavon G, Samek W, Müller K. Methods for interpreting and understanding deep neural networks. Digit Signal Process. 2018;73:1-15. doi:10.1016/j.dsp.2017.10.011

10. Maier O, Menze BH, von der Gablentz J, et al. ISLES 2015 - A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med Image Anal. 2017;35:250-269. doi:10.1016/j.media.2016.07.009

11. Kistler M, Bonaretti S, Pfahrer M, Niklaus R, Büchler P. The virtual skeleton database: an open access repository for biomedical research and collaboration. J Med Internet Res. 2013;15(11):e245. Published 2013 Nov 12. doi:10.2196/jmir.2930

12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. Proc IEEE Int Conf Comput Vis. 2017;2017-Octob:618-626. doi:10.1109/ICCV.2017.74

13. Jezzard P, Barnett AS, Pierpaoli C. Characterization of and correction for eddy current artifacts in echo planar diffusion imaging. Magn Reson Med. 1998;39(5):801-812. doi:10.1002/mrm.1910390518

Figures

Model Hyperparameter	Value
Learning rate (tuned)	0.001
Batch size (tuned)	8
Dropout rate (tuned)	0.5
L2 regularization rate (tuned)	0.001
Epochs	10
Loss function	binary cross-entropy
Optimizer	Adam
Activation function	ReLU
Output activation	Sigmoid
Number of filters	8/16
Filter size	(3 x 3 x 3)
Pooling size	(2 x 2 x 2)

Figure 1. Model hyperparameters used during training. The table shows the selected hyperparameters for the final model. Batch size, learning rate, l2 normalization rate and dropout rate were fine-tuned using 5-fold cross validation with grid search.

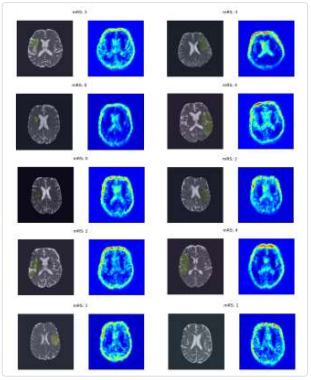

Figure 2. Illustration of class activation maps for correctly predicted patients from the training set. The figure shows, for each patient, a slice from the ADC map where the lesion is most visible with the lesion mask overlayed. The generated heatmap corresponding to the same slice is shown beside the original image. On the heatmaps, red areas indicate high interest, followed by yellow areas. The mRS score of each patient is given above their maps.

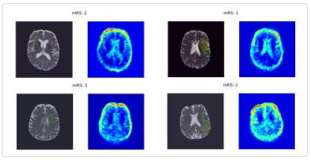

Figure 3. Illustration of class activation maps for correctly predicted patients from the validation set. The figure shows, for each patient, a slice from the ADC map where the lesion is most visible with the lesion mask overlayed. The generated heatmap corresponding to the same slice is shown beside the original image. On the heatmaps, red areas indicate high interest, followed by yellow areas. The mRS score of each patient is given above their maps.