Automatic Quality Control of Human Brain T1w MRI Scans

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neuroimaging studies often involve running complex software pipelines on large imaging datasets. Each image must undergo quality control (QC) before analysis because of a number of possible artifacts that can compromise its quality. No gold standard exists for QC, whether manual or automatic, because studies have different aims, and the threshold of acceptable image quality may vary depending on the research interest. Manual QC is a tedious and long process, and while many good protocols exist, humans exhibit high inter- and intra-rater variability, which makes QC a good candidate for automation. In this work, deep convolutional neural network (CNN) models are presented that automatically predict QC ratings. For the Infant Brain Imaging Study (IBIS [1]), sensitivity of 97% and specificity of 96% were achieved predicting a single imaging expert's rating, but the model was unable to generalize to data from different sources. To address the inability to generalize, a second tri-planar CNN model was trained on the Autism Brain Imaging Data Exchange (ABIDE [2]) dataset to predict two raters' QC ratings.

1 Introduction

Quality Control is a subjective and challenging problem for neuroimaging studies. QC protocols vary depending on the downstream application, for instance if the investigation is only interested in a particular brain region. Additionally, different imaging experts/software pipelines may have different thresholds for what they consider an acceptable image. If a T1w image fails quality control, a new scanning session must be booked or the subject is dropped from the study. An automatic within-session system with minimal processing requirements could be deployed as a tool for MRI technicians to re-run scans that have a high probability of failure, avoiding costly scanner re-bookings and study delays.

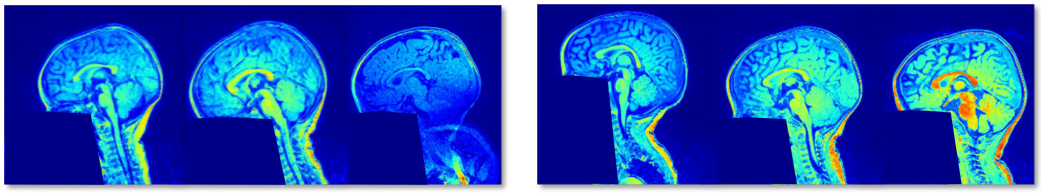


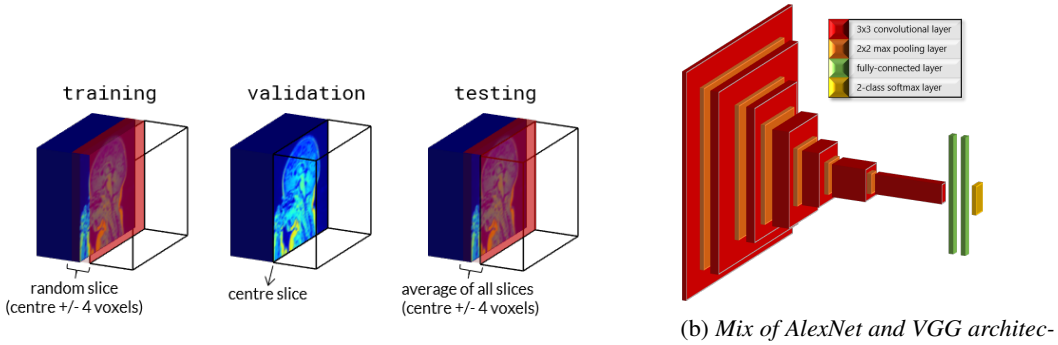
Figure 1: *Examples of images that failed QC (left) and passed QC (right) as determined by a single rater. The failures are due to subject motion and scanning artifacts. Face information was removed from these images to prevent identification of subjects.*

Medical images can be corrupted by subject motion, artifacts induced by improper scanner parameters, or magnetic susceptibility effects created by subject anatomy or foreign objects in the scanner. Some images simply show poor contrast between anatomically significant regions, or are very noisy. Fig. 1 shows examples of images that have passed and failed QC. Several attempts have been made to automate QC of T1w MRI. The MRIQC package (`https://github.com/poldracklab/mriqc`) extracts a variety of Image Quality Metrics (IQM [3]) and uses a supervised classifier to predict manual QC ratings based on two raters trained on the same QC protocol with 76% accuracy [4].

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

2 Experiments

This work poses automatic QC as a supervised learning problem, where ground truth QC labels come from an imaging expert for a large set of MRI scans. Two similar experiments have been performed on two separate datasets. The first dataset is a subset (N=1163) of the IBIS longitudinal study, where participants are infants aged 6 months to 2 years. This dataset is comparatively homogeneous; all scans were acquired using the same MRI protocol, with the same parameters, and many scans are corrupted by motion due to the difficulty in scanning an infant-based cohort. A single imaging expert, whose intra-rater reliability is unknown, produces the QC labels using LORIS, a databasing system for managing large studies (`https://loris.ca/`) [5]. A CNN with structure shown in Fig. 2 is trained to learn QC labels, with 70% of subjects used for training and 15% for each of the validation and testing sets. 97% sensitivity and 96% specificity was achieved, and the proposed 2D CNN outperformed both a CNN with 3D convolutions and fully-connected neural networks.



(a) *Slice sampling strategy*

(b) *Mix of AlexNet and VGG architectures [6, 7].*

Figure 2: *CNN inputs (left) and structure (right).*

The second dataset is a combination of ABIDE-I [2] (N=1102), which combines MRI from 17 sites with no common protocol, and the UCLA Consortium for Neuropsychiatric Phenomics LA5c Study (OpenFMRI accession number ds000030) [8] (N=265). These images must be registered with an affine transformation to MNI-space because they are not all the same size/resolution. QC labels for the dataset were produced by two raters at the Stanford Centre for Reproducible Neuroscience and demonstrated to have "fair to moderate" inter-rater reliability. This is a more challenging dataset because of the extreme variability in terms of the age of participants (7-64 years old), scanner manufacturer, and MRI protocol differences. For this experiment, slices are extracted from the three orthogonal planes of the MRI scan, and slices are treated as different input channels to the CNN. The three inputs each go into a sub-network consisting of four convolutional layers, which are concatenated and trained end to end. This architecture, while redundant, is necessary to capture three-dimensional relationships in the image that are not possible to capture using 3D convolutions because the number of parameters required to learn exceeds the capacity of the powerful systems available for this work.

Preliminary experiments using a similar training/validation split suggest accuracy higher than the 76% reported by the MRIQC classifier is achievable, with one experiment reaching 78% accuracy on the validation set, but we have yet to show that our model generalizes to the held-out test set.

3 Conclusion

Learning a reproducible, fully-automatic QC system has been demonstrated. While it remains unclear whether the QC protocol policy learned by the systems presented in this work are optimal for all downstream applications, CNNs are well-suited to this type of image recognition task and have the potential to replace time-consuming and subjective QC that is currently only possible by highly-trained experts. Future plans to extend this work include adding guided Grad-CAM visualizations to identify which parts of the image caused the QC failure [9] and automating QC of derivatives, such as automatically-extracted gray/white matter surfaces.

4 Acknowledgements

This work would not have been possible without the imaging data collected as part of the Infant Brain Imaging Study. The Infant Brain Imaging Study Network is comprised of the following individuals: J. Piven, H.C. Hazlett, C. Chappell, S.R. Dager, A.M. Estes, D. Shaw, K.N. Botteron, R.C. McKinstry, J. Constantino, J.R. Pruett, R.T. Schultz, S. Paterson, L. Zwaigenbaum, J.T. Elison, A.C. Evans, D.L. Collins, G.B. Pike, V. Fonov, P. Kostopoulos, S. Das, G. Gerig, M.A. Styner, and H. Gu.

We thank the Infant Brain Imaging Study children and families for their ongoing participation in this longitudinal study.

References

[1] Lewis, J. D., Evans, A. C., Pruett, J. R., Botteron, K. N., McKinstry, R. C., Zwaigenbaum, L., Estes, A.M., Collins, D.L., Kostopoulos, P., Gerig, G., Dager, S.R., Paterson, S., Schultz, R., Styner, M.A., Hazlett, H.C., Piven, J. (2017). The Emergence of Network Inefficiencies in Infants With Autism Spectrum Disorder. Biological psychiatry.

[2] Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., ... & Deen, B. (2014). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Molecular psychiatry, 19(6), 659-667.

[3] Shehzad, Z., Giavasis, S., Li, Q., Benhajali, Y., Yan, C., Yang, Z., ... & Craddock, C. (2015). The Preprocessed Connectomes Project Quality Assessment Protocol—a resource for measuring the quality of MRI data. Frontiers in neuroscience.

[4] Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., & Gorgolewski, K. J. (2017). MRIQC: Advancing the automatic prediction of image quality in MRI from unseen sites. PloS one, 12(9), e0184661.

[5] Das, S., Zijdenbos, A. P., Vins, D., Harlap, J., & Evans, A. C. (2012). LORIS: a web-based data management system for multi-center studies. Frontiers in neuroinformatics, 5, 37.

[6] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

[7] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[8] Poldrack, R. A., Congdon, E., Triplett, W., Gorgolewski, K. J., Karlsgodt, K. H., Mumford, J. A., ... & Bilder, R. M. (2016). A phenome-wide examination of neural and cognitive function. Scientific data, 3, 160110.

[9] Selvaraju, R. R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., & Batra, D. (2016). Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. arXiv preprint arXiv:1610.02391.