Medical Image Quality Assurance using Deep Learning

Dženan Zukić[1] DZENAN.ZUKIC@KITWARE.COM
Anne Haley[1] ANNE.HALEY@KITWARE.COM
Curtis Lisle[2] CLISLE@KNOWLEDGEVIS.COM
Kilian Pohl[3] KILIAN.POHL@STANFORD.EDU
Hans J. Johnson[4] HANS-JOHNSON@UIOWA.EDU
Aashish Chaudhary[1] AASHISH.CHAUDHARY@KITWARE.COM
[1] *Kitware Inc., Carrboro, North Carolina, USA*
[2] *KnowledgeVis LLC, Florida, USA*
[3] *Department of Psychiatry & Behavioral Sciences, Stanford University*
[4] *Electrical and Computer Engineering, University of Iowa*

Editors: Under Review for MIDL 2022

Abstract

We present an open-source web tool for quality control of distributed imaging studies. To minimize the amount of human time and attention spent reviewing the images, we created a neural network which provides an automatic assessment. This steers reviewers' attention to potentially problematic cases, reducing the likelihood to miss an image quality issue. We test our approach using 5-fold cross validation on a set of 5217 magnetic resonance images.
Keywords: quality control, quality assurance, neural networks, web interface.

1. Introduction

Discovery of new knowledge in medicine is sometimes accomplished by large, multi-center imaging studies. The success of these studies depends on the quality of images and the resulting measurements, regardless of the size of the study. Our open-source Medical Image Quality Assurance (MIQA) system represents a design that facilitates collaboration and sharing. It incorporates a state-of-the-art deep learning component to improve the effectiveness of Quality Control (QC) efforts unique to the needs of multi-center studies. The usefulness of this unique QC system is being tested by National Consortium on Alcohol and Neurodevelopment in Adolescence (NCANDA) project. NCANDA uses magnetic resonance images (MRIs) of the brain, so we focus on deep learning using head MRIs in this paper.

MIQA is a client-server web application based on Girder and django. It uses Vue and Vuetify for graphical user interface (GUI). For image processing we use Insight Toolkit (ITK). For neural network related operations we use PyTorch, MONAI, and TorchIO.

Images are auto-assessed after upload. Images which are not reviewed are available in a queue to tier one reviewers. They can mark it as good or questionable. Questionable images need to be reviewed by tier two reviewers who then make a final decision. An image can be marked bad only if it has presence of at least one artifact. This can be indicated in the GUI, or a free-form text comment provided if the problem does not match well the pre-defined artifact classes.

As far as we know, this is the first attempt at assessing image quality of 3D images. Previous studies used photographs (Bosse et al., 2017; Hosu et al., 2020), retinal fundus images (Yu et al., 2017), or tried to improve image quality (Higaki et al., 2019). The closest one focuses on quantifying motion artifact (Butskova et al., 2021), but that is only one of nine artifacts we consider here. Their ultimate goal was to correct motion artifacts.

2. Materials and Methods

We use data from PREDICT-HD study (Paulsen et al., 2014), which has manually assessed quality for structural (T_1, T_2, PD) brain MRIs. In this study we used only the T_1-weighted images. 2299 were acquired on 1.5T and 2918 on 3T MRI machines. The most important annotation is overall quality, scored on 0-10 scale (see Figure 1). It also has manual assessment of signal to noise ratio and contrast to noise ratio. We didn't use them as they are highly correlated the with overall quality, with Pearson coefficients of 0.721 and 0.715.

There was binary presence indication of anatomical variants and nine artifacts for most images. Some of the images were missing a few of the indications. Of the 5217 T_1 images, 520 had normal anatomical variants, 61 had lesions, 269 incomplete brain coverage, 30 misalignment, 28 had wraparound, 347 ghosting, 585 inhomgeneity, 187 metal susceptibility, 888 flow artifact, and 1286 had truncation.

The data mostly consisted of images with little or no artifacts. 392 images had quality five or lower (considered bad by PREDICT-HD experimenters), while 4825 had quality six or higher. We augmented training data to compensate for this class imbalance. For this, we extended the TorchIO library and applied random augmentations, each with probability ranging from 10% to 50%. We implemented five simulated artifacts: ghosting, motion, inhomogeneity, spike, and noise.

Since the images have variable size (in our case ranging from 192x256x104 to 512x512x256), we decided to split them into tiles of size 64^3 with minimal overlap. We apply the neural network (NN) to each tile, and then average the outputs. NN has 5 convolutional layers and a fully connected layer with eleven outputs. In the loss function, we combine regression to overall quality with the focal loss for each of the ten presence indicators. For training we use AdamW optimizer and exponential learning rate schedule. We trained for preset number of epochs, determined experimentally to allow convergence.

3. Results and Conclusion

We split the available data into 5 folds based on subject identifier. We used coefficient of determination R^2 applied to the overall quality (0-10) to assess predictive power. On validation data, R^2 was: 0.33, 0.27, 0.33, 0.24 and 0.14 for the five folds. Compare that to training data where it was: 0.65, 0.57, 0.64, 0.54 and 0.67. As there are no previous studies to compare to, we are setting precedent.

We provide source code at `https://github.com/OpenImaging/miqa`, where we openly develop this system. We also maintain a demo instance at `https://miqa.miqaweb.io/`. Next steps are: adding permutation and swapping of axes as an additional augmentation for training, applying this to all 9475 structural images, and transferring this to NCANDA data (which is not defaced).

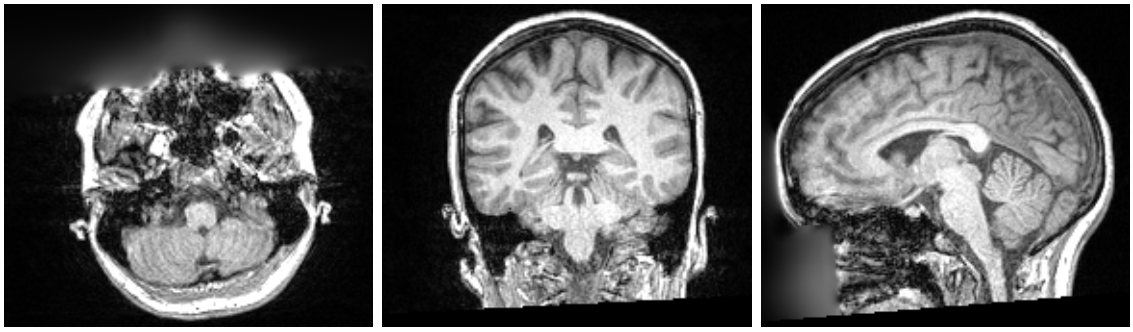

Figure 1: Slices of a T_1 weighted image with an overall score of 8 and no artifacts present. All the images from PREDICT-HD set are defaced with a facial blur.

Acknowledgments

We would like to thank our colleagues: Scott Wittenburg, Daniel Chiquito, Zach Mullen, Matt McCormick, Jeff Baumes (all at Kitware) and James Klo at Stanford.

References

Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing*, 27(1):206–219, 2017.

Anastasia Butskova, Rain Juhl, Dženan Zukić, Aashish Chaudhary, Kilian M Pohl, and Qingyu Zhao. Adversarial bayesian optimization for quantifying motion artifact within mri. In *Int. Workshop on PRedictive Intellig. In MEdicine*, pages 83–92. Springer, 2021.

Toru Higaki, Yuko Nakamura, Fuminari Tatsugami, Takeshi Nakaura, and Kazuo Awai. Improvement of image quality at ct and mri using deep learning. *Japanese journal of radiology*, 37(1):73–80, 2019.

Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020.

Jane S Paulsen, Jeffrey D Long, Hans J Johnson, Elizabeth H Aylward, Christopher A Ross, Janet K Williams, Martha A Nance, Cheryl J Erwin, Holly K Westervelt, Deborah Lynn Harrington, et al. Clinical and biomarker changes in premanifest huntington disease show trial feasibility: a decade of the predict-hd study. *Front. in aging neuroscience*, 6:78, 2014.

FengLi Yu, Jing Sun, Annan Li, Jun Cheng, Cheng Wan, and Jiang Liu. Image quality classification for dr screening using deep learning. In *2017 39th Annual Int. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 664–667. IEEE, 2017.